Current Research in Food Science 9 (2024) 100867

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Unveiling intraspecific diversity and evolutionary dynamics of the foodborne pathogen *Bacillus paranthracis* through high-quality pan-genome analysis

Yuhui Du [a,1], Chengqian Qian [b,c,1], Xianxin Li [c,1], Xinqian Zheng [a], Shoucong Huang [d], Zhiqiu Yin [e,*], Tingjian Chen [a,**], Li Pan [b,***]

[a] *MOE International Joint Research Laboratory on Synthetic Biology and Medicines, School of Biology and Biological slaEngineering, South China University of Technology, Guangzhou, 510006, Guangdong, PR China*
[b] *School of Biology and Biological Engineering, Guangzhou Higher Education Mega Centre, South China University of Technology, Guangzhou, 510006, Guangdong, PR China*
[c] *Foshan Branch of Tianyan (Tianjin) High-tech Co., Ltd, Foshan, 528000, Guangdong, PR China*
[d] *Foshan Haitian (Gaoming) Flavouring Food Co., Ltd, Foshan, 52a8000, Guangdong, PR China*
[e] *Department of Clinical Laboratory, Key Laboratory of Biological Targeting Diagnosis, Therapy and Rehabilitation of Guangdong Higher Education Institutes, The Fifth Affiliated Hospital, Guangzhou Medical University, Guangzhou, 510700, Guangdong, PR China*

ARTICLE INFO

Handling Editor: Professor A.G. Marangoni

Keywords:
Bacillus paranthracis
Foodborne pathogen
Adaptation
Evolutionary dynamics
Gene loss
Emerging antimicrobial resistance

ABSTRACT

Understanding the evolutionary dynamics of foodborne pathogens throughout host-associated habitats is of utmost importance. Bacterial pan-genomes, as dynamic entities, are strongly influenced by ecological lifestyles. As a phenotypically diverse species in the *Bacillus cereus* group, *Bacillus paranthracis* is recognized as an emerging foodborne pathogen and a probiotic simultaneously. This poorly understood species is a suitable study model for adaptive pan-genome evolution. In this study, we determined the biogeographic distribution, abundance, genetic diversity, and genotypic profiles of key genetic elements of *B. paranthracis*. Metagenomic read recruitment analyses demonstrated that *B. paranthracis* members are globally distributed and abundant in host-associated habitats. A high-quality pan-genome of *B. paranthracis* was subsequently constructed to analyze the evolutionary dynamics involved in ecological adaptation comprehensively. The open pan-genome indicated a flexible gene repertoire with extensive genetic diversity. Significant divergences in the phylogenetic relationships, functional enrichment, and degree of selective pressure between the different components demonstrated different evolutionary dynamics between the core and accessory genomes driven by ecological forces. Purifying selection and gene loss are the main signatures of evolutionary dynamics in *B. paranthracis* pan-genome. The plasticity of the accessory genome is characterized by horizontal gene transfer (HGT), massive gene losses, and weak purifying or positive selection, which might contribute to niche-specific adaptation. In contrast, although the core genome dominantly undergoes purifying selection, its association with HGT and positively selected mutations indicates its potential role in ecological diversification. Furthermore, host fitness-related dynamics are characterized by the loss of secondary metabolite biosynthesis gene clusters (BGCs) and CAZyme-encoding genes and the acquisition of antimicrobial resistance (AMR) and virulence genes via HGT. This study offers a case study of pan-genome evolution to investigate the ecological adaptations reflected by biogeographical characteristics, thereby advancing the understanding of intraspecific diversity and evolutionary dynamics of foodborne pathogens.

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: yzq7873728@126.com (Z. Yin), chentj@scut.edu.cn (T. Chen), btlipan@scut.edu.cn (L. Pan).
[1] Yuhui Du, Chengqian Qian, and Xianxin Li contributed equally to this paper as first authors.

https://doi.org/10.1016/j.crfs.2024.100867
Received 8 September 2024; Received in revised form 20 September 2024; Accepted 20 September 2024
Available online 21 September 2024

1. Introduction

Bacteria, which live in widely diverse niches, have adapted to various ecosystems in parallel with their genetic evolution, leading to extraordinary genomic and phenotypic diversification (Boutte and Crosson, 2013)(Anani et al., 2020). Microorganisms existing in many habitats are considered generalists, whereas those restricted to a single environment are specialists (Thomas et al., 2016)(Malard et al., 2019). The concept of the bacterial pan-genome, which represents the entire gene set across all strains of a species, was first proposed in 2005 (Tettelin et al., 2005). Ecological selection pressures drive the adaptive evolution of the pan-genome through gene gain and loss. Adaptive gene gains via horizontal gene transfer (HGT) generate significant evolutionary novelty (Arnold et al., 2022)(Power et al., 2021), whereas gene losses have been observed to arise as specific adaptations to the symbiotic lifestyle (Pande et al., 2014). As a result, there is remarkable variation in the gene content of individual genomes, resulting in fascinating genetic differences in their pan-genomes. These differences, which are crucial for adaptive differentiation within bacterial species, are reflected in various features of the pan-genome (Maistrenko et al., 2020)(Hartmann and Croll, 2017), including its state (open versus closed), the size of each component, and the functional enrichment of components.

The rapid increase in genome data has spurred a corresponding rise in efforts to understand the mechanistic, ecological, and evolutionary forces that shape bacterial pan-genomes (Brockhurst et al., 2019). Currently, pan-genome analysis, as a whole, suffers from several known issues, including limited data, genomic fragmentation, incompleteness and contamination, and confounding and redundant (high similarity) strains. These issues can lead to biases and errors in the results of pan-genome analysis and further mislead the downstream analysis. The inclusion of fragmented and incomplete genomes leads to significant core gene loss, and contaminated sequences significantly affect the identification of accessory genes (Li and Yin, 2022). Confounding strains can adversely affect the identification of core genes, and highly similar strains can skew the results of pan-genome diversity and the identification of strain-specific genes (Wu et al., 2021)(Yang and Gao, 2022). Therefore, constructing a high-quality pan-genome on the basis of a representative dataset enhances analysis efficiency and ultimately contributes to a deeper understanding of the evolutionary dynamics of a bacterial species (Wu et al., 2022).

The *Bacillus cereus* group, more broadly known as *B. cereus* sensu lato (s.l.), is a species complex of environmentally ubiquitous spore-forming gram-positive bacteria (Y. Liu et al., 2017). It comprises at least 21 closely related proposed species/genomospecies, ranging from pathogens to probiotics (Carroll et al., 2022)(Carroll et al., 2020), including the bioterrorism agent *B. anthracis*, the foodborne pathogen *B. cereus*, and the biopesticide *B. thuringiensis*. The *B. cereus* group is a suitable study model for microbial evolutionary ecology because it has adapted and radiated to exploit environmental niches (Raymond and Bonsall, 2013). However, information on the adaptive evolution of the pan-genome among species within the *B. cereus* group is limited. *Bacillus paranthracis*, a representative species in the *B. cereus* group, was first described in 2017 (Y. Liu et al., 2017). This species is recognized as an emerging opportunistic human pathogen (Matson et al., 2020)(Carroll et al., 2019), a plant growth-promoting bacteria, and a probiotic (Bukharin et al., 2019) simultaneously. Members of this species have been isolated from various environments, including human samples, soil, and the rhizosphere. The diverse lifestyles and predictable genetic diversity of this species make it suitable for analyzing the link between pan-genome evolution and host-associated adaptation.

To date, owing to the limited studies on *B. paranthracis*, very little molecular data on this species have been presented. Consequently, we performed a case study to understand intraspecific diversity and evolutionary dynamics in *B. paranthracis* driven by ecological adaptation through high-quality pan-genome analysis. As of May 2023, more than 150 *B. paranthracis* genomes are available in the National Center for Biotechnology Information (NCBI) GenBank database. However, because the taxonomic classification of the *B. cereus* group is notoriously convoluted, many *Bacillus* sp. genomes have not been assigned to species-level taxonomic units. The EzBioCloud database has deposited more than 130 additional whole-genome assemblies identified as *B. paranthracis* (Yoon et al., 2017). Owing to these nearly 300 genomes, in-depth studies of *B. paranthracis* based on a high-quality pan-genome are now possible. Here, by combining pan-genome and comparative genomic analyses, we aim to expand our understanding of bacterial pan-genome evolution driven by ecological adaptation. Metagenomic read recruitment analyses were performed to explore the biogeographical characteristics of *B. paranthracis*. To further our understanding of the evolutionary dynamics of the pan-genome, we comparatively analyzed the functional enrichment and selective pressure of the core genome and accessory genomes. Genomic plasticity was assessed through analysis of HGT and gene gain and loss. We also investigated the genetic basis of key properties within the pan-genome, such as secondary metabolism, carbohydrate-active enzymes (CAZymes), and genotypic and phenotypic profiles related to virulence and antimicrobial resistance (AMR), to elucidate the mechanisms of ecological adaptation in *B. paranthracis*.

2. Materials and methods

2.1. Biogeographic distribution analysis of B. paranthracis

We queried the representative 16S rRNA sequence of *B. paranthracis* (accession: KJ812420; 1509 bp) against 500,048 Sequence Read Archives (SRAs) (accessed on Jun. 7, 2023) via the Integrated Microbial Next Generation Sequencing (IMNGS) platform (Lagkouvardos et al., 2016) with a minimum DNA size of 200 bp and a threshold of 99%. An SRA sample was included in our study if it contained at least three reads that mapped to the query 16S rRNA sequence (Table S1).

2.2. Genome collection and filtering

A comprehensive search for *B. paranthracis* genomes were obtained from the taxonomically united genome database in EzBioCloud (Yoon et al., 2017) and NCBI GenBank. The taxonomic framework and species members of the *B. cereus* group are described in Laura's review (Carroll et al., 2022). All *Bacillus* sp. genomes were downloaded from the NCBI GenBank database, which was accessed in May. 2023. Detailed information on each genome is listed in Table S2. The initial collection encompassed 290 genomes, comprising 269 *B. paranthracis* genomes and 21 reference genomes representing other species within the *B. cereus* group. The quality assessment of these genomes was conducted via CheckM v1.0.13 (Parks et al., 2015). We subsequently excluded genomes with an excessive number of contigs (>300), less than 95% completeness, or more than 10% contamination. Following these stringent filters, our final collection consisted of 242 *B. paranthracis* genomes (Table S3), complemented by 21 reference genomes of closely related species. To ensure unified gene finding and reannotation, we utilized Prokka v1.14.5 software (Seemann, 2014) for all the genomes in our collection.

2.3. Pan-genome analysis

Orthologous groups of protein families within the pan-genome were delimited via the OrthoFinder2 software with the DIAMOND method (Emms and Kelly, 2015)(Buchfink et al., 2015). Curve fitting for the pan-genome was conducted via power-law regression in accordance with Heap's law ($n = \kappa N^{\gamma}$) (Tettelin et al., 2008; Heaps, 1978), where N is the number of genomes, κ is the proportionality constant, and the growth exponent $\gamma > 0$ suggests an open pan-genome.

2.4. Phylogenetic analysis

The phylogenetic analysis of the core genome was conducted via single nucleotide polymorphisms (SNPs) across single-copy core gene families, as extracted from the OrthoFinder output files. The nucleotide sequences of these gene families were extracted based on their protein accession numbers and aligned via MAFFT v7.508 software (Katoh and Standley, 2013). To avoid phylogenetic confusion, we identified and excluded putative recombination regions from the SNPs via ClonalFrameML v1.12 software (Didelot and Wilson, 2015). Finally, the maximum likelihood (ML) tree was constructed via MEGA 11 (Tamura et al., 2021) with the general time reversible (GTR) model and 100 bootstrap replicates.

Using this Manhattan distance matrix based on the binary presence or absence of each pan-genome gene family within the individual genomes, we constructed a pan-genome tree with MEGA 11, employing the neighbor-joining (NJ) method (Tamura et al., 2021). To assess the congruence between the core and pan-genome trees, we calculated the normalized Robinso–Foulds (nRF) scores via TreeCmp (Bogdanowicz et al., 2012). A comparative analysis of these two trees was conducted via the Dendroscope 3 program (Huson and Scornavacca, 2012). Additionally, a binary matrix representing the presence or absence of each pan-genome gene family across all the genomes was generated for constructing a network phylogeny using the Neighbor-Net algorithm implemented in SplitsTree5 software (Bagci et al., 2021).

2.5. Population structure analysis

The population genetic structure of *B. paranthracis* was inferred via STRUCTURE 2.3.4 (Falush et al., 2007) based on SNPs derived from single-copy core gene families, employing 20,000 burn-in cycles, 20,000 sampling MCMC cycles, K values ranging from 2 to 12, and five independent replicates. The optimal K value, indicative of the most likely number of genetic clusters, was identified via STRUCTURE Harvester (Earl and vonHoldt, 2012).

2.6. Comparative genomic analysis

The average nucleotide identities (ANI) were calculated via fastANI v2.0 (Jain et al., 2018). Gene clusters associated with secondary metabolism were identified and characterized via antiSMASH 6.1.1 (Medema et al., 2011) with the default parameters. Homology analysis of these biosynthetic gene clusters (BGCs) was conducted via BiG-SCAPE v1.1.5 software (Navarro-Muñoz et al., 2020). CAZyme-coding genes were identified through a search of the dbCAN2 database (H. Zhang et al., 2018) with HMMER (Finn et al., 2011). The functional annotation of pan-genome gene families was performed based on the Clusters of Orthologous Groups (COG) categories (Galperin et al., 2015) via eggNOG-mapper 2.1.9 software (Huerta-Cepas et al., 2017). AMR and virulence genes were detected with Abricate v1.0.1 (https://github.com/tseemann/abricate), utilizing the CARD database and Virulence Factors Database (VFDB) (Alcock et al., 2020)(B. Liu et al., 2022). Plasmid nucleotide sequences were distinguished from assembled contigs or scaffolds via the GPU Docker image-based deeplasmid (Andreopoulos et al., 2022). Virulence factors were also identified by aligning all protein sequences with BLASTp against the Pathogen Host Interactions database (PHI-base 5.0 with an E-value cutoff $<$ 1e-6, an identity $>$60%, and a coverage $>$60%) and BTyper 3.0 (Urban et al., 2020)(Carroll et al., 2020). Macromolecular system detection was carried out via MacSyFinder v2 (Touchon et al., 2014) and TXSScan v1.1.1 (Abby and Rocha, 2017). The gene family gains and losses at each node and branch within the 80 *B. paranthracis* genomes were inferred via CAFÉ 5 (Mendes et al., 2021) with default parameters. HGTector v2.0 (Zhu et al., 2014) with the database 2021-11-21 was used to identify potential horizontally transferred genes, employing the *B. cereus* group (Rank: species group; Taxon ID: 86661) and *Bacillus* (Rank: genus; Taxon ID: 1386) as the self-group and close-group, respectively.

2.7. Selection pressure analysis

Selection pressure in coding regions is estimated by calculating the ratio of the nonsynonymous to the synonymous substitution rates, denoted as *dN/dS*. ParaAT 2.0 software was used for codon-based alignment of orthologous genes (Z. Zhang et al., 2012). The Fast Unconstrained Bayesian Approximation (FUBAR) pipeline (Murrell et al., 2013) integrated within HYPHY v2.5.42 software was subsequently employed to calculate the *dN/dS* values across each site within the orthologous gene families.

3. Results and discussion

3.1. Biogeographic distribution and abundance profiles indicate that B. paranthracis is a generalist species with diverse habitats, thriving in host-associated habitats

To explore the biogeographical distribution of *B. paranthracis*, we searched the representative 16S rRNA sequence (accession: KJ812420) against 500,048 public SRA datasets via the IMNGS platform, accessed on 7 Jun 2023 (Lagkouvardos et al., 2016). We identified 29,759 amplicon samples matching the *B. paranthracis* 16S rRNA sequence by at least three reads matching, which were distributed across 14 habitat categories as detailed in Table S1. Mapping of these samples revealed a widespread distribution of *B. paranthracis* (Fig. 1A). Significant associations were observed with samples from "Terrestrial" (n = 9832; 33.0%), "Plants" (n = 5970; 20.1%), "Human" (n = 4098; 13.8%), "Mammals" (n = 3929; 13.2%), and "Aquatic" (n = 2932; 9.85%) habitats (Fig. 1B). This pattern suggests that *B. paranthracis* has been isolated from a wide range of environments, spanning both host-associated and non-host associated habitats. Moreover, the strains included in our genome dataset were sourced from a variety of samples, including soil, food, plant, sediment, sewage, seawater, and human sampples (Table S3). Meijenfeldt et al. recently defined the social niche breadth (SNB) score as a quantitative measure of the microbial niche range (von Meijenfeldt et al., 2023). According to this study, the genus *Bacillus* qualifies as a social generalist with a high SNB score of 0.448, which is evident in its prevalence across 4974 out of 22,518 samples (von Meijenfeldt et al., 2023). Consequently, *B. paranthracis* can be considered a niche generalist.

The distribution of *B. paranthracis* varies significantly across different habitats. *B. paranthracis* presented a relatively high average relative abundance in samples associated with the "Mammals" (1.512 $\pm$ 5.556%) and "Human" (0.899 $\pm$ 5.568%) habitats, in contrast to the relatively low average relative abundance found in the "Aquatic" (0.436 $\pm$ 2.614%) and "Plants" (0.496 $\pm$ 3.051%) habitats (Fig. 1C). The "Terrestrial"-associated samples presented an average relative abundance of *B. paranthracis* of 0.630 $\pm$ 3.816%. We further screened 2464 samples with high abundances that presented $\geq$1% relative abundance values. As expected, *B. paranthracis* was more prevalent in "Mammals" (n = 691; 28.0%) and "Human" (n = 404; 16.4%) associated samples (Fig. S1), accounting for nearly half of the highly abundant samples. The high abundance in these samples suggests that *B. paranthracis* members can locally outcompete their neighbors, indicating adaptability and competitiveness in corresponding habitats (von Meijenfeldt et al., 2023). The presence and relative abundance of *B. paranthracis* indicate a preference for host-associated habitats. Indeed, this species contains a considerable number of clinical isolates (Table S3) and is considered to constitute a species hosting an overweight of human pathogenic strains.

3.2. Selection of high-quality representative genomes

To determine the evolutionary relationships among these genomes, we constructed a phylogenetic tree based on SNPs in 1501 single-copy

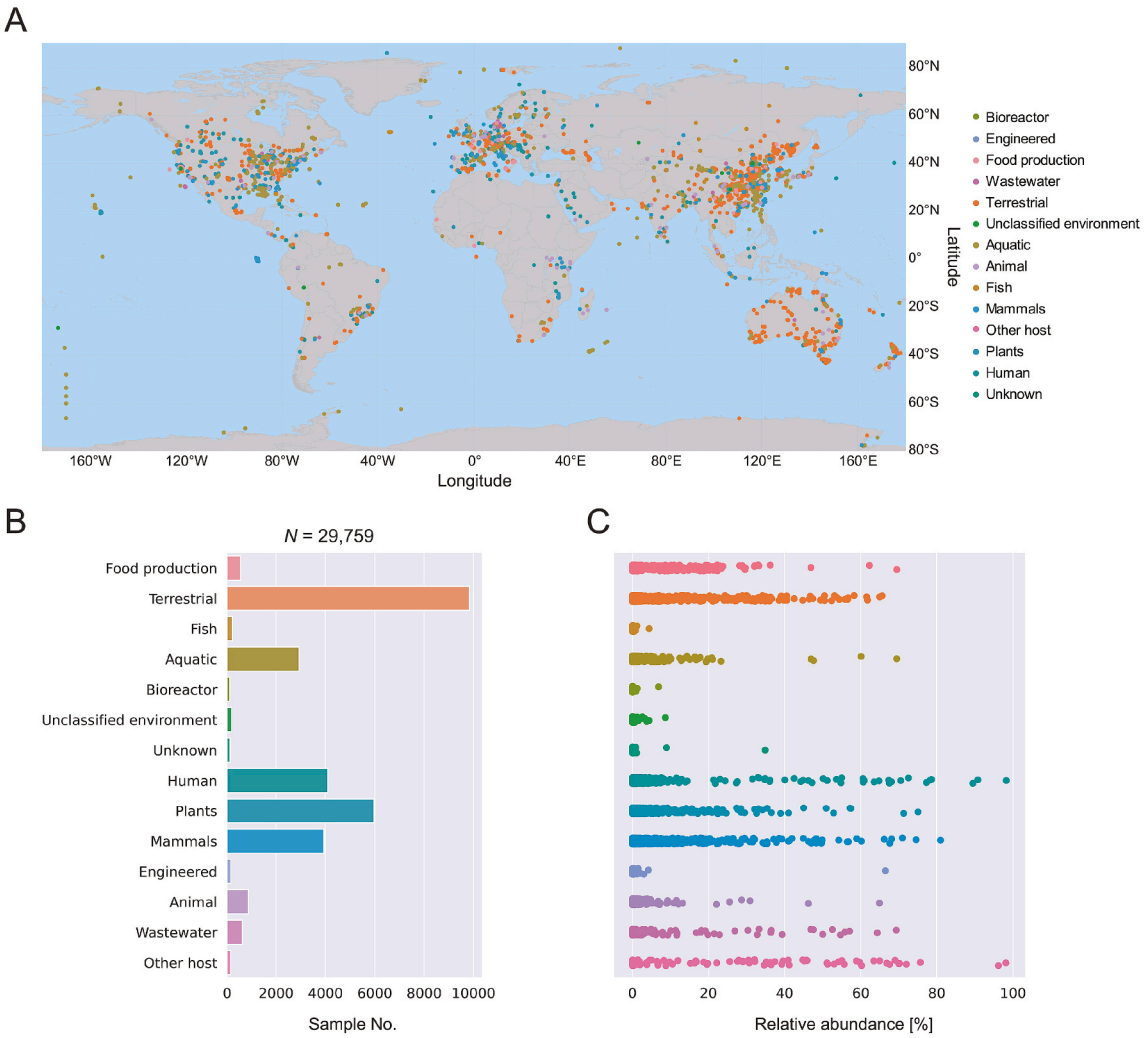

Fig. 1. Biogeographic distribution and relative abundance profiles of *B. paranthracis*. A. Global distribution of *B. paranthracis*. The map displays colored circles representing the biogeographical origins of the collected samples, each isolated from distinct environments. Only amplicon samples with *B. paranthracis* 16S rRNA sequences exhibiting ≥99% identity and at least three mapped reads were included in this analysis. The color scheme corresponds to the environmental categories from which the samples were obtained. B. Environmental percentage of amplicon samples from various environments. C. Relative abundance of *B. paranthracis* 16S rRNA sequences within different environmental categories.

core gene families, which were shared by 242 high-quality *B. paranthracis* genomes (Fig. S2) and 21 reference genomes of other proposed species within the *B. cereus* group (Carroll et al., 2022). The core genome tree (Fig. 2A) revealed that the closest species to *B. paranthracis* is *Bacillus pacificus* (strain anQ-h4). The *B. paranthracis* members formed a monophyletic clade and then fell into four major intraspecific genetic populations based on the STRUCTURE analysis (Fig. 2A). The ANI values derived from comparisons between *B. paranthracis* and other *B. cereus* group species ranged from 80.3 ± 0.1% (*B. manliponensis*) to 96.0 ± 0.1% (*B. pacificus*), aligning well with the proposed 96% threshold for species delineation within the *B. cereus* group (Y. Liu et al., 2017) (Fig. 2A). *B. paranthracis* members presented high ANI values, averaging 98.2 ± 1.0% among themselves (Fig. 2A). Based on these findings, we selected 242 validated *B. paranthracis* genomes for further genomic analysis. On average, *B. paranthracis* members have a genome size of 5472.8 ± 235.3 kb, comprising 5568.2 ± 271.6 protein-coding genes, 79.6 ± 29.7 tRNAs, and 13.8 ± 12.4 rRNAs (Fig. 2B). The variations in genome components among these *B. paranthracis* genomes reflect a degree of genetic heterogeneity. The GC content of the *B. paranthracis* genome showed minor variation, with an average value of 35.3 ± 0.002%. Geographically, these strains were isolated across 21 countries globally between 1972 and 2022, predominantly in the last decade, representing geographic and temporal diversities (Fig. 2C, Fig. S3A, and Table S3).

However, our dataset includes numerous genomes with a high degree of genomic similarity. For example, many members of population 3 share high ANI values and form a clade with exceedingly short branch lengths. The presence of these redundant strains could introduce bias in pan-genome analysis (Chan et al., 2015)(Du et al., 2023), particularly for determining strain-specific gene content (Wu et al., 2021; Yang and Gao, 2022). To mitigate this bias, we constructed a high-quality pan-genome for *B. paranthracis* by selecting representative genomes and excluding redundant genomes with ANI values > 99.8%. As a result, we retained a total of 80 high-quality representative genomes, which maintained the majority of the geographic and temporal diversity (Fig. S3B and Fig. S3C).

3.3. *High-quality B. paranthracis pan-genome reveals extensive genetic diversity and differential functional enrichment*

The representative dataset resulted in a high-quality pan-genome of *B. paranthracis*, comprising 14,095 homologous gene families (Fig. 3A and Table S4). Among these, 3645 found in all the genomes constitute the core genome, accounting for 25.9% of the pan-genome; 6094 present

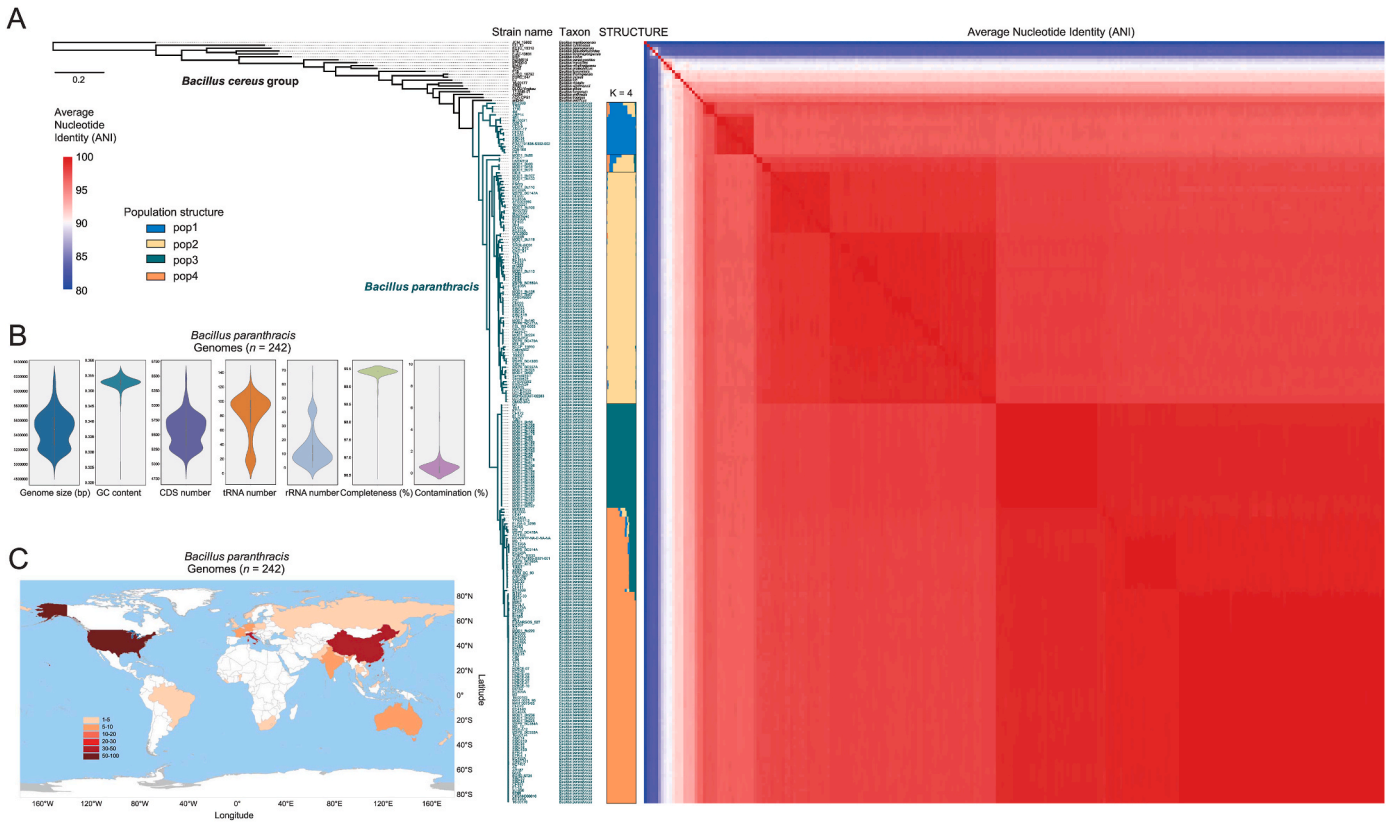

Fig. 2. Genetic relatedness of *B. paranthracis* among the *Bacillus cereus* group. A. Core genome phylogeny. The maximum likelihood (ML) tree was constructed via single-nucleotide polymorphisms (SNPs) across 1501 single-copy core gene families shared by 263 *Bacillus* sp. genomes. Adjacent to the tree, colored blocks represent the genetic clusters inferred via STRUCTURE, whereas the accompanying heatmap displays pairwise average nucleotide identities (ANI). B. Overview of the genomic characteristics of 242 *B. paranthracis* genomes. The violin plots illustrate the distribution of genomic features, with centerlines representing the medians, violin edges marking the 25th and 75th percentiles, and whiskers extending to 1.5 times the interquartile range. C. Geographic distribution map of 242 *B. paranthracis* strains. The colored blocks reflect the number of strains isolated from different countries and regions.

at least two but not all *B. paranthracis* genomes make up the accessory genome, forming the largest portion at 43.2% of the pan-genome; and the remaining 4356 genes (30.9%), which are unique to individual genomes, represent the strain-specific gene content (Fig. 3A). In the core genome, 3122 represent single-copy core gene families, with the remaining 523 existing in multiple copies. In the individual genomes, the accessory gene families varied in size from 1222 to 2326, averaging 1781.2 ± 207.1 gene families. The strain-specific genes range from 11 to 336 genes, with an average of 54.5 ± 54.3 genes. The variable gene contents of the genome compositions exhibit a high level of divergence. Most variable (non-core) gene families are not widely distributed ($\leq$2 genomes), which is likely facilitated by the exclusion of redundant genomes (Yang and Gao, 2022) (Fig. S4A). The core and pan-genome curves revealed a steady increase in pan-genome size with each additional genome, whereas the core genome size decreased (Fig. 3B). Heaps' power law function ($n = \kappa N^{\gamma}$) is used to determine whether a pan-genome is open ($\gamma \geq 0$) or closed ($\gamma < 0$) (Tettelin et al., 2008). The growth exponent value (γ) for the pan-genome curve is 0.253, indicating that *B. paranthracis* has an open pan-genome. Notably, the exclusion of strain-specific genes results in a plateau in the pan-genome accumulation curve, suggesting that most undiscovered genes occur in individual genomes.

To elucidate the role of variable gene families in pan-genome architecture dynamics, we constructed and compared phylogenetic trees for both the core and pan-genome. The core genome tree was generated using SNPs from 3122 single-copy core gene families, and the pan-genome tree was constructed based on the presence/absence of non-core gene families. As shown in Fig. 3C, the phylogenetic positions of more than half of the *B. paranthracis* strains were congruent between the two trees. However, the remaining strains presented discordances in their phylogenetic positions and branching orders. The discordances are also quantified by a high nRF score of 0.720, where a score closer to 0 (ranging from 0 to 1) suggests greater congruence between the two trees (Bogdanowicz et al., 2012). Additionally, the Neighbor-Net pan-genome tree revealed a reticular network indicative of substantial homologous recombination and HGT (Fig. S4B). This obscured some aspects of phylogenetic inertia (core genome tree: Fig. 3C), and led to a variable non-core gene repertoire. The microbial genome size is known to vary and is correlated with different habitats (von Meijenfeldt et al., 2023). Consistent with this, *B. paranthracis*, as a generalist in habitats with high local diversity, presents an open pan-genome with notable genetic diversity, particularly within individual genomes.

We compared the predicted biological functions of each component in the pan-genome using the COG annotation. Owing to the limited functional studies, 5787 (41.1%) gene families lacked COG functional annotation, categorized as "HP: Hypothetical proteins", which are predominantly found in the accessory genome and strain-specific gene content (Fig. 4A). Core genome enrichment was observed for genes related to "J: Translation, ribosomal structure and biogenesis" as well as metabolism categories such as energy production and conversion, amino acid, nucleotide, coenzyme, and inorganic ion metabolism [COG-J, -C, -E, -H, and -P: Fisher's exact test, P-value <0.01; COG-F: Fisher's exact test, P-value <0.05]. These essential genes enable *B. paranthracis* to efficiently take up nutrients from the environment and maintain a basic lifestyle. The accessory genome is mainly responsible for transcription, replication, recombination, repair, defense mechanisms, and cell wall functions [COG-K, -L, and -M: Fisher's exact test, P-value <0.01; COG-V: Fisher's exact test, P-value <0.05]. The strain-specific genes are notably

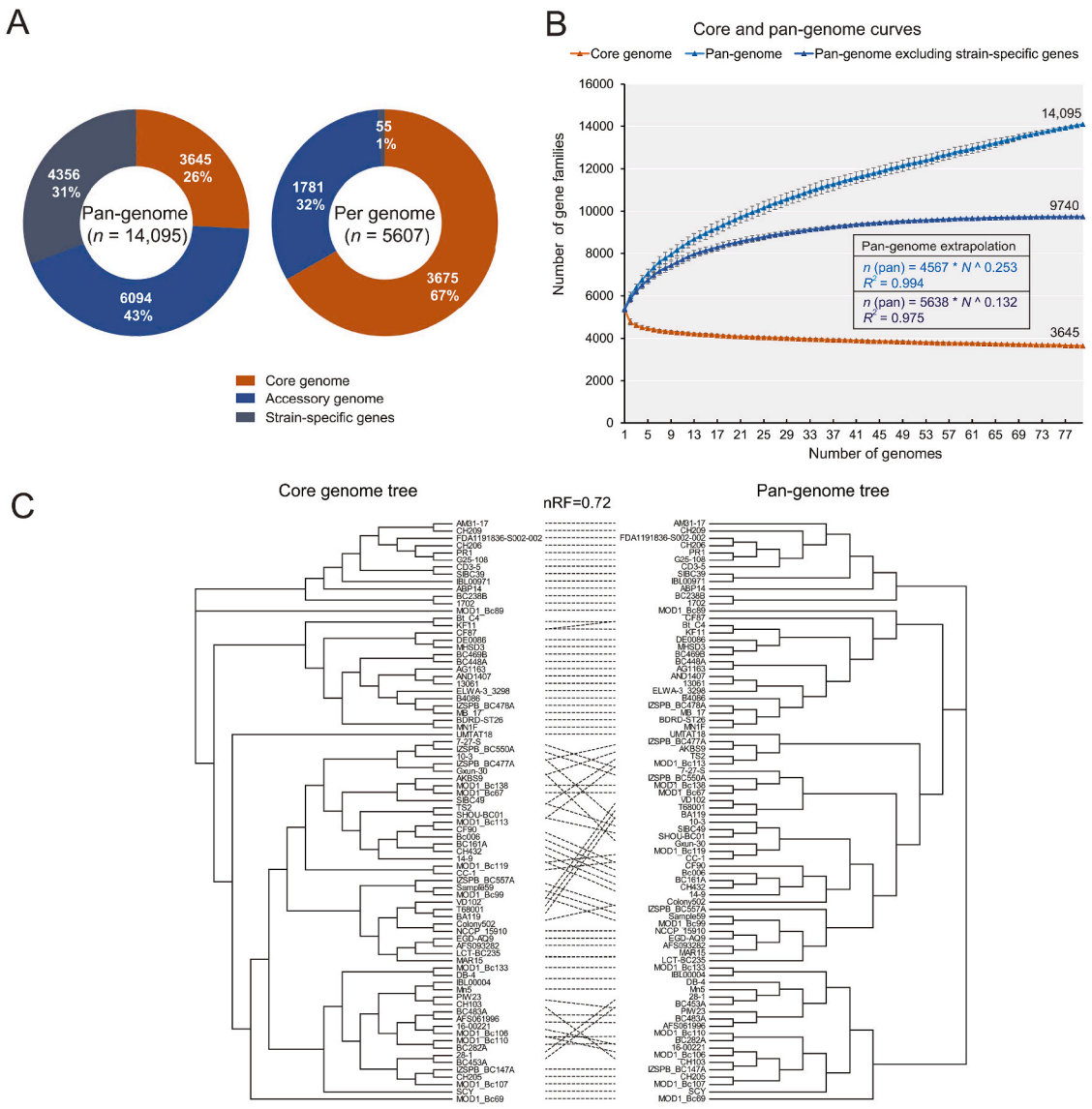

Fig. 3. Pan-genome analysis of *B. paranthracis* utilizing 80 representative genomes. A. Proportional distribution of pan-genome components. Pie charts depict the percentage compositions of the core, accessory, and strain-specific gene families within both the pan-genome and individual genomes. B. Progressive curves for the core genome, pan-genome, and pan-genome excluding strain-specific genes. The core gene families trended to decrease with the addition of genomes, whereas the pan-gene families tended to increase. The inferred mathematical functions describing the pan-genome curves are shown in the graph. C. Comparison of phylogenetic trees generated via single-copy core gene families and the pan-genomic distance metric. Congruence was measured via normalized Robinson–Foulds (nRF) scores.

involved in "L: Replication, recombination and repair" [Fisher's exact test, *P*-value <0.05] and "S: Function unknown" [Fisher's exact test, *P*-value <0.01]. Both the accessory genome and strain-specific genes are enriched in gene families associated with replication, recombination, and DNA repair, underscoring the mechanisms that sustain an open pan-genome. This finding also highlights the significant role of genetic recombination in the evolutionary dynamics of the *B. paranthracis* pan-genome. Considering the importance of the non-core genome for ecological adaptation (Azarian et al., 2020), individually distributed genes of unknown function may contribute to the ecological diversification of *B. paranthracis* strains, and their biological functions warrant further investigation.

3.4. Distinct purifying selection across pan-genome components and a low-cost evolutionary strategy: prevalent positively selected mutations in the core genome

Evolutionary dynamics for bacterial adaptation are dominated by natural selection, as many bacterial species have large effective population sizes (N_e), which increases the efficacy of selection (Arnold et al., 2022). We calculated the *dN/dS* via a codon-level analysis of natural selection across 7944 orthologous families, encompassing 3645 core and 4299 accessory gene families present in at least four genomes. Most gene families presented low *dN/dS* values, with an average of 0.214 ± 0.228, indicating predominant purifying or stabilizing selection in the *B. paranthracis* pan-genome. Selective signatures differ between the core and the accessory genomes. The core gene families presented a significantly lower average *dN/dS* value of 0.149 ± 0.132 than the accessory gene families (0.274 ± 0.276) [*t*-test, $P < 0.01$], highlighting the distinct degree of purifying selection between the core and accessory genomes (Fig. 4B). Significantly stronger purifying constraints were observed for the core gene families across most functional categories and HP, with the exception of the COG-H "Coenzyme transport and metabolism" (Fig. 4C). A total of 60 gene families presented *dN/dS* values exceeding one, indicating positive selection. Most of them ($n = 50$) are accessory gene families, whereas only three are core gene families; the functions of

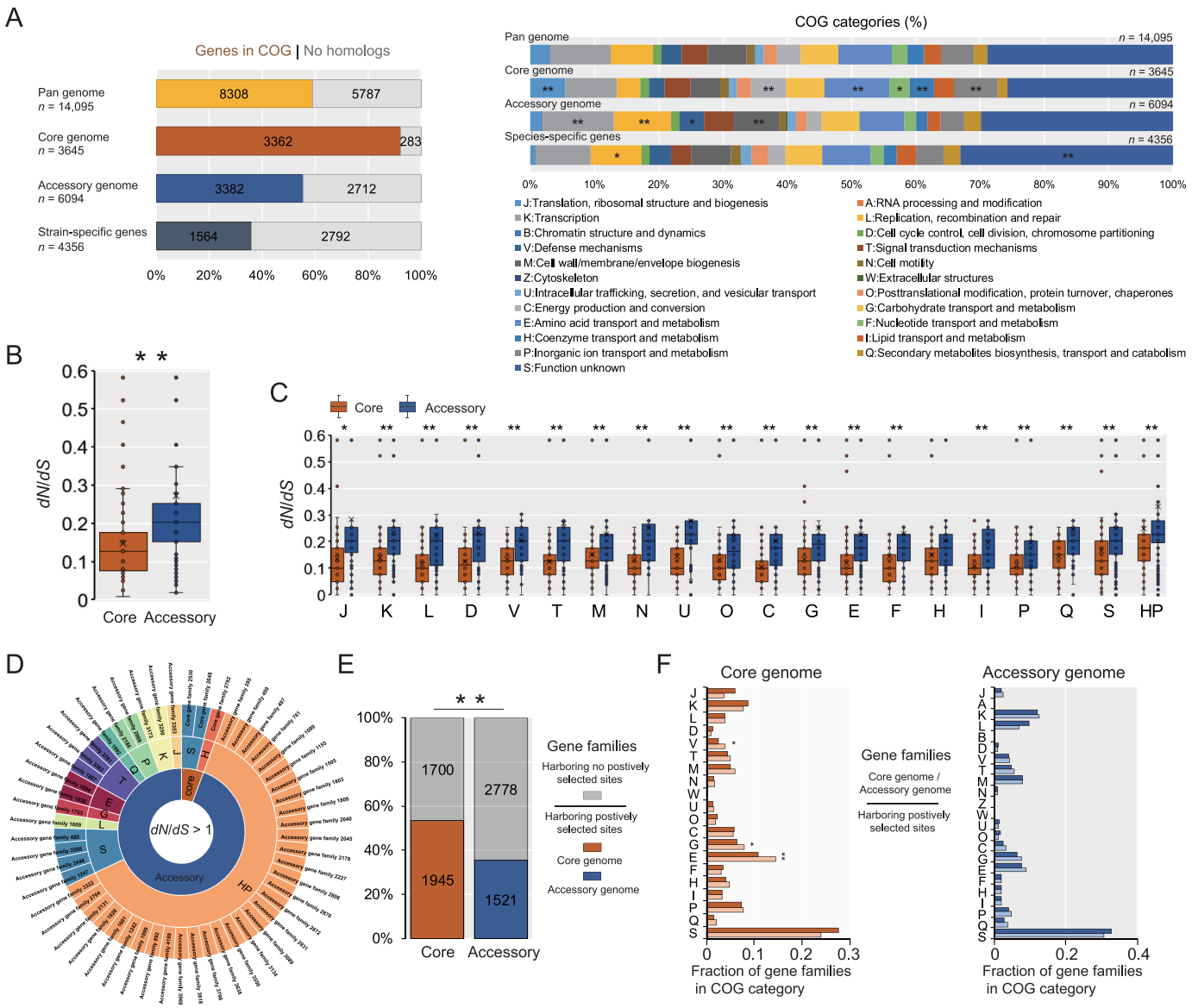

Fig. 4. Comparative analysis of functional enrichment and selective pressures within the *B. paranthracis* pan-genome components. A. Distribution of Clusters of Orthologous Groups (COG) functional categories for pan-genome components. * for Fisher's exact test, *P*-value <0.05; ** for Fisher's exact test, *P*-value <0.01. B. Comparisons of the nonsynonymous/synonymous rate (*dN/dS*) ratios of gene families between the core and accessory genomes. ** for *t*-test, *P*-value <0.01. C. Comparisons of the *dN/dS* values of gene families in COG functional categories between the core and accessory genomes. * for *t*-test, *P*-value <0.05; ** for *t*-test, *P*-value <0.01. D. Sunburst diagram of positively selected gene families (*dN/dS* > 1), categorized by COG functional categories and pan-genome components. The inner ring represents pan-genome components, the middle ring represents COG functional categories, and the outer ring represents positively selected gene families. E. Diagram of the relationships of gene families harboring positively selected sites between the core and accessory genomes. ** for chi-squared test, *P*-value <0.01. F. Distribution of COG functional categories for gene families harboring positively selected sites in the core and accessory genomes. * for Fisher's exact test, *P*-value <0.05; ** for Fisher's exact test, *P*-value <0.01.

the majority of these gene families remain unknown (COG-S and HP) (Fig. 4D). As expected, the core genome is more conserved than the accessory genome because of stronger purifying selection. Compared with the corresponding core genomes, microbial accessory genomes, which harbor more mobile genes, are subjected to vastly different selective pressures (Castillo-Ramírez et al., 2011; Bohlin et al., 2017). Some accessory gene families exhibit high *dN/dS* values due to positive selection, potentially undergoing adaptive changes as they are not yet at the peak of their fitness landscape for niche specialization (Azarian et al., 2020).

Despite the entire coding regions of gene families being under purifying selection, we observed numerous codon sites with significant evidence of positive selection (posterior probability ≥0.9) within pan-gene families. A total of 3466 gene families were found to contain one or more such codon sites. Among these, 1945 (56.1%) are core gene families, whereas the remaining 1521 (43.5%) are accessory gene families. Interestingly, even though the entire coding region has undergone stronger purifying selection, core gene families are more likely to harbor positively selected sites than accessory gene families (1945 out of 3645 versus 1521 out of 4299) [Chi-squared test, $\chi^2 = 258.6$, df = 1, $p < 0.0001$] (Fig. 4E). This finding implies that conserved proteins have undergone evolutionary modifications at the residue level. The resulting functional shifts in core genes might provide new ecological opportunities, potentially triggering phenotypic modifications or niche partitioning. Furthermore, core gene families with positively selected sites were significantly associated with COG-V, -E, and -G [COG-E: Fisher's exact test, *P*-value <0.01; COG-V and -G: Fisher's exact test, *P*-value <0.05] (Fig. 4F). Indeed, adaptive changes in metabolism categories

(COG-E and -G) are typically correlated with the colonization of diverse ecological niches (Goyal, 2018)(Cummins et al., 2022).

Generally, purifying selection on core gene families, the backbones of bacterial genomes, is crucial for maintaining the basic functions of a species. In contrast, flexible gene content is primarily implicated in niche-specific adaptations (Brockhurst et al., 2019)(Chattopadhyay et al., 2009). However, the costs of HGT, such as genome disruption, cytotoxic effects, energy costs, and the risk of disrupting intracellular interactions, cannot be ignored, particularly in the core genome (Baltrus, 2013). Thus, it can be inferred that a low-cost/compensatory strategy is employed by *B. paranthracis*, which could influence its ecological diversification and fitness. In this strategy, mutations potentially driven by positive selection fine-tune the core properties, particularly their metabolic capabilities.

3.5. HGT from distant Bacillaceae relatives reflects B. paranthracis adaptation to terrestrial habitats

HGT is a fundamental part of bacterial evolution, providing raw material for natural selection (Arnold et al., 2022)(Koonin and Wolf, 2008). In the *B. paranthracis* pan-genome, we identified 790 horizontally transferred gene families originating from distant taxonomic groups, including 342 (43.3%; 275.9 ± 4.7 per genome) core, 356 (45.1%; 110.2 ± 13.8 per genome) accessory, and 92 (11.7%; 1.2 ± 2.0 per genome) strain-specific genes (Fig. 5A). Notably, nearly half of these gene families represent core genome. High rates of genetic exchange within the core genome are recognized as crucial for the adaptive evolution of bacteria with global populations in response to selection pressures (Everitt et al., 2014)(Preska Steinberg et al., 2022). Thus, these gene families might contribute to the core properties of *B. paranthracis* for adaptive evolution. Furthermore, these gene families experienced significantly stronger purifying selection, with average *dN/dS* values of 0.155 ± 0.126, which are lower than the pan-genome-wide average *dN/dS* value of 0.214 ± 0.228 [*t*-test, $P < 0.01$] (Fig. 5B). Among these, stronger evolutionary constraints are observed in the horizontally transferred core gene families (average $dN/dS = 0.128 \pm 0.085$) than in the accessory genome (average $dN/dS = 0.188 \pm 0.158$) (Fig. 5B). Recently transferred genes often evolve rapidly and are quickly lost if

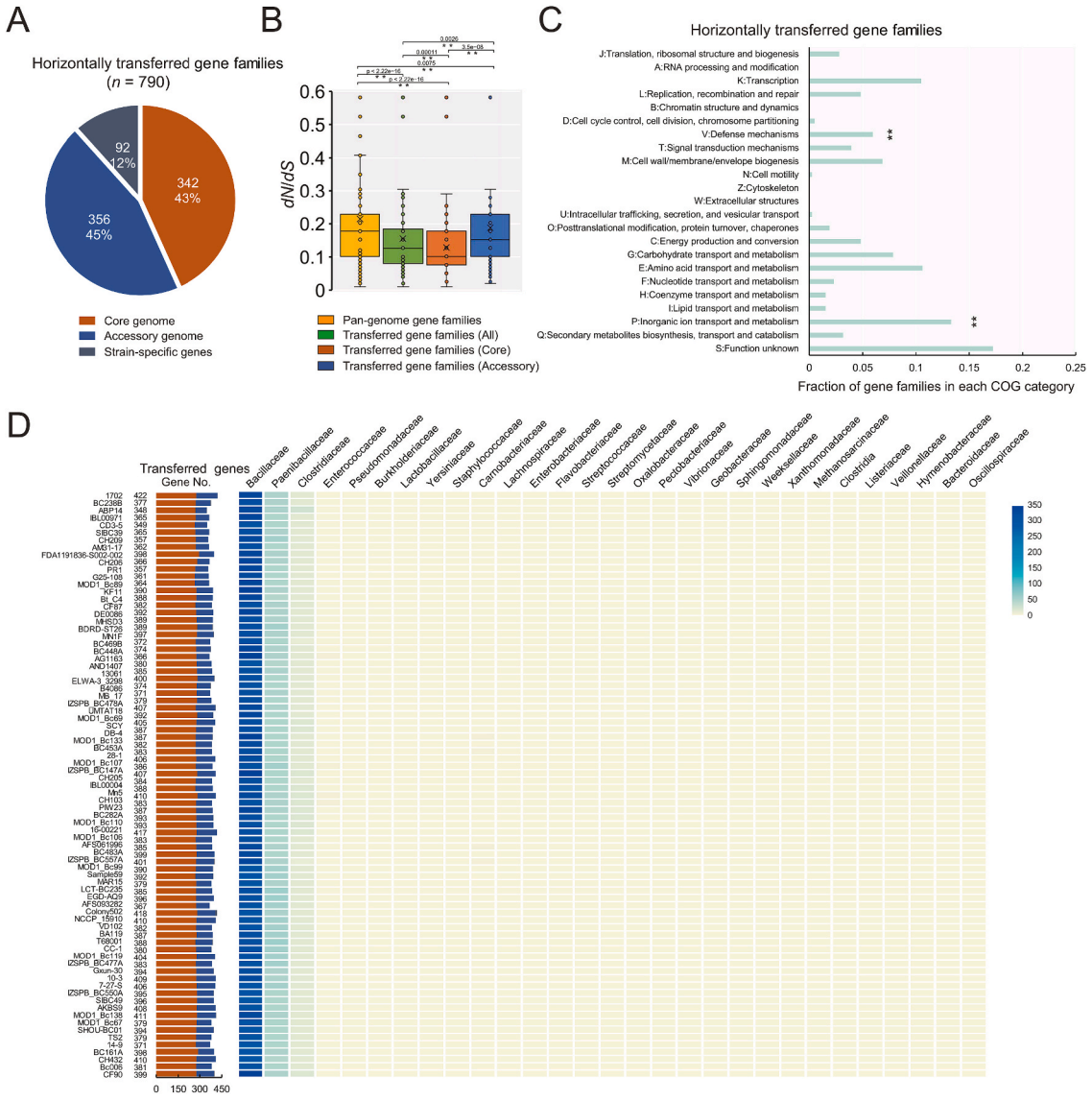

Fig. 5. Horizontal gene transfers (HGTs) in the *B. paranthracis* pan-genome. A. Pan-genome component distribution of potential horizontally transferred gene families. B. Pairwise *dN/dS* value comparisons for horizontally transferred gene families among pan-genome components. ** for *t*-test, *P*-value <0.01. C. Functional categorization of potential horizontally transferred gene families. ** for Fisher's exact test, *P*-value <0.01. D. Distribution of horizontally transferred genes acquired in each strain and the potential donor bacterial taxa implicated in HGT events.

their fitness is reduced. Conversely, if they offer selective advantages for niche adaptation, they may be retained for longer periods (Hao and Golding, 2006). The dominant purifying selection acting on the horizontally transferred gene families in the *B. paranthracis* pan-genome may indicate that these acquired genes have already been purged with fitness-decreasing mutations before being transferred into *B. paranthracis* (Castillo-Ramírez et al., 2011). The retained genes that have undergone stronger purifying selection might play important roles in adaptive evolution and speciation. For example, in rhizobial bacteria, HGT and purifying selection appear to be particularly strong in genes associated with their symbiosis, which is beneficial for their adaptation to the host environment (Epstein and Tiffin, 2021).

The horizontally transferred gene families were significantly enriched in COG-V (n = 47; 5.9%) and -P (n = 105; 13.3%) [Fisher's exact test, P-value <0.01] (Fig. 5C). Genes within COG-V (defense mechanisms) have been reported to be predominant in bacterial genomic islands, which are frequently mediated by HGT (Merkl, 2006). The COG-P (inorganic ion transport and metabolism) category is thought to influence bacterial fitness in soil and liquid environments (Morales et al., 2023). Biogeographic characterization indicated that *B. paranthracis* is widely distributed across terrestrial and aquatic samples (Fig. 1A). Ecology is a principal determinant of HGT (Smillie et al., 2011), with gene transfer limited by ecological opportunity and the habitats occupied by the recipient species. Therefore, *B. paranthracis* has an increased capacity for inorganic ion transport and metabolism through HGT. A total of 29 potential HGT donor bacterial families were identified, with Bacillaceae (311.4 ± 13.7) and Paenibacillaceae (49.2 ± 3.1) being the most common (Fig. 5D). Gene transfer requires close physical proximity of microbes, so HGT is likely most efficient between immediate neighbor isolates from the same niche location (Babic et al., 2011). Bacterial habitat preferences are phylogenetically predetermined (Konstantinidis and Tiedje, 2005), which means that bacterial species from the same genus or multiple closely related genera tend to share more genes. Since *B. paranthracis* belongs to the Bacillaceae (*Bacillus* genus), the co-occurrence of *B. paranthracis* with distant members from other genera of Bacillaceae and Paenibacillaceae in the same habitats increases the efficiency of HGT.

3.6. Gene loss as the dominant evolutionary process shaping *B. paranthracis* pan-genome evolution

To gain insight into the evolutionary dynamics of *B. paranthracis* pan-genome evolution, we determined the gained and lost gene families for each evolutionary node and branch on the core genome tree. The

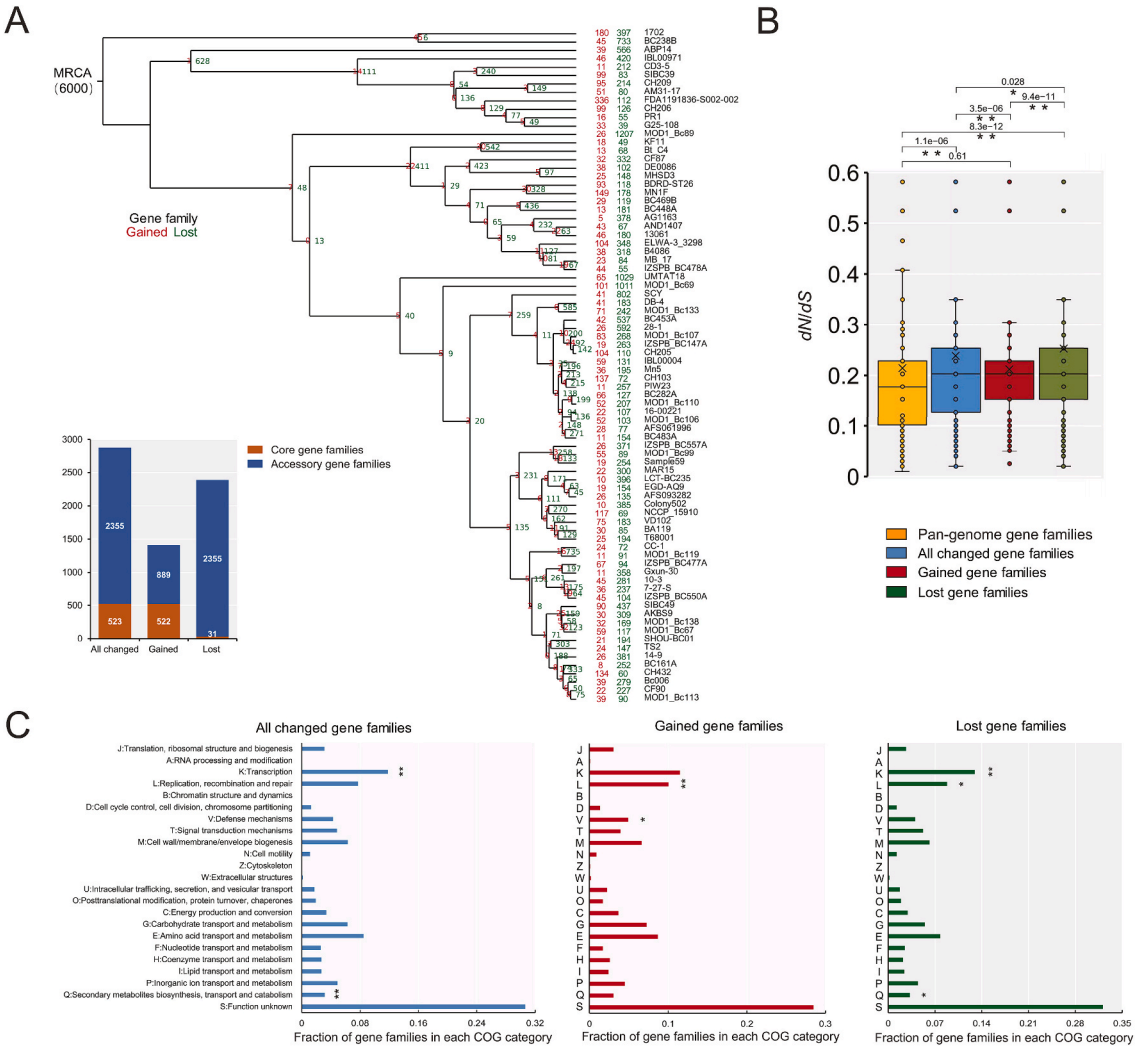

Fig. 6. Evolutionary dynamics of gene families within the *B. paranthracis* pan-genome. A. Gene gain and loss along each evolutionary branch of the core genome tree. Red and green indicate the numbers of gained and lost gene families, respectively, for each node. A bar chart summarizing the core, accessory, and strain-specific gene families among those that have changed, been gained, or lost is provided. B. Pairwise *dN/dS* value comparisons for the entire pan-genome and for gene families that have changed, been gained, or lost. * for *t*-test, P-value <0.05; ** for *t*-test, P-value <0.01. C. Functional categorization of changed, gained, and lost gene families. * for Fisher's exact test, P-value <0.05; ** for Fisher's exact test, P-value <0.01.

genome of the most recent common ancestor (MRCA) of *B. paranthracis* is inferred to contain 6000 protein-coding genes, which exceeds the average of 5568.2 ± 271.6 per genome among extant species. This suggests a significant reduction in genome size over time. We identified 2878 changed/evolving gene families that experienced gain and/or loss events, including 2386 lost and 1441 gained gene families (Fig. 6A). Notably, 522 core gene families experienced gain events, accounting for nearly all multiple copies in the core genome (522 out of 523). On average, each node/branch on the phylogenetic tree presented 30.1 ± 41.0 gains and 208.4 ± 196.5 losses, indicating that gene loss is the dominant evolutionary process for *B. paranthracis* [*t*-test, $P < 0.01$] (Fig. S5). Furthermore, we observed a greater frequency of gene gain and loss events at the tips of the tree than at the tips of the internal branches, highlighting individual-specific evolution (Fig. 6A). The loss of dispensable genes in free-living organisms can provide a selective advantage by conserving limiting resources in the microbial community (Morris et al., 2012). Overall, we infer that *B. paranthracis* individuals have adopted a genome streamlining strategy to adapt to diverse environmental conditions.

In the *B. paranthracis* pan-genome, these changed gene families generally presented significantly weaker evolutionary constraints, as indicated by higher average *dN/dS* values of 0.239 ± 0.214 than the pan-genome-wide average of 0.214 ± 0.228 [*t*-test, $P < 0.01$] (Fig. 6B). Notably, the lost gene families presented significant signatures of weaker evolutionary constraints, with average *dN/dS* values of 0.253 ± 0.227 [*t*-test, $P < 0.01$], whereas the gained gene families did not (average $dN/dS = 0.212 \pm 0.145$; [*t*-test, $P = 0.61$]) (Fig. 6B). Positive selection was observed in 20 lost gene families, in contrast to only one gained gene family. Generally, relaxed purifying selection and positive selection can lead to the accumulation of deleterious mutations, resulting in ongoing losses of genes and functions. Natural selection appears to be the primary driver of adaptive gene loss in the *B. paranthracis* pan-genome. The prevailing view is that increased mutation rates can increase adaptive capacity but may also lead to gene loss, thereby promoting genome reduction in prokaryotes (Taddei et al., 1997)(Bourguignon et al., 2020). Consequently, *B. paranthracis* may shed costly, dispensable genes in leaky environments to increase fitness.

3.7. Functional enrichment of gained and lost gene families reveals functional changes for adaptive evolution

The functional properties of the changed gene families were significantly associated with COG-K and -Q [Fisher's exact test, *P*-value <0.01]. The gained gene families were significantly enriched in COG-L [Fisher's exact test, *P*-value <0.01] and -V [Fisher's exact test, *P*-value <0.05] (Fig. 6C). Moreover, we detected significant enrichment of positively selected sites and HGT events within the COG-V category. These findings are consistent with those of Rasigade et al.'s study, which revealed that recombination and positive selection signatures across genes are involved in defense mechanisms, especially those for AMR, indicating the host adaptation of pathogenic *B. cereus* group members (Rasigade et al., 2018).

Gene losses were significantly enriched in COG-K [Fisher's exact test, *P*-value <0.01], -L, and -Q [Fisher's exact test, *P*-value <0.05] (Fig. 6C), indicating a reduced requirement for these biological functions. The expression of dispensable genes associated with core cellular functions such as transcription (COG-K) and translation can be highly toxic (Sorek et al., 2007)(Szabová et al., 2011), making them likely candidates for rapid loss (Brockhurst et al., 2019). As a result, the enrichment of COG-K within the accessory genome of *B. paranthracis* and other microorganisms (Fig. 4A) (Zhong et al., 2018; Zhong et al., 2019) reflected an adaptive response of core cellular functions to diverse environments. The gene families related to DNA replication, recombination, and repair (COG-L) were significantly enriched in both the gained and the lost gene families. The variability of these genes is likely crucial for the adaptive plasticity of the *B. paranthracis* genome, allowing different isolates to adjust their recombination and mutation rates. Specific ensembles of BGCs related to secondary metabolites in soil bacteria often reflect environmental factors more closely than phylogenetic factors do because of niche-driven selective pressure on BGC retention (Sharrar et al., 2020). Genomic changes in BGCs through gene losses within COG-Q may indicate the adaptation of *B. paranthracis* individuals to specialized niches. The following sections explore the genetic repertoires associated with key biological processes in the *B. paranthracis* pan-genome, providing in-depth insights into the dynamics of adaptive evolution.

3.8. Gene loss and dispersal of secondary metabolite BGCs in shaping host-associated adaptation

Many soil microbes produce secondary metabolites through BGCs, playing crucial ecological roles in their complex and heterogeneous microenvironments (Sharrar et al., 2020). In the *B. paranthracis* pan-genome, 199 putative BGCs across six classes were identified, with an average of 2.488 ± 3.543 BGCs per genome (Table S5 and Fig. 7A). These BGCs were further classified into 42 homologous families via BiG-SCAPE software (Navarro-Muñoz et al., 2020). RiPPs and non-ribosomal peptide synthetases (NRPSs) were the most abundant, with 95 and 43 BGCs across 23 and eight families, respectively (Fig. 7B). The secondary metabolites produced by *Bacillus* spp. Are valuable for biocontrol and plant growth promotion (Shen et al., 2023), indicating the great potential of *B. paranthracis* isolates in agricultural and biotechnological applications. All the BGC families, as non-core genetic elements, were scattered throughout the *B. paranthracis* genomes (Fig. 7A), with more than 85% (36 out of 42) of the families present in fewer than 10 strains. Nine strains harbored more than eight BGCs, whereas more than half of the strains ($n = 52$) lacked any detectable BGCs. These results indicate a low level of conservation and significant individual variation in secondary metabolite biosynthesis among *B. paranthracis* strains. Ecological forces driving gene duplications, HGT, gene loss and shuffling in bacteria lead to variations in the abundances of secondary metabolite biosynthesis gene clusters in the prokaryotic genome (Seshadri et al., 2022). In the *B. paranthracis* pan-genome, most BGC families (39 out of 42) presented signatures of gene gain and loss, with 262 gene families having experienced loss events and 218 having experienced gain events. This result is consistent with previous findings that gene loss predominates in *B. paranthracis* pan-genome evolution, particularly regarding the substantial loss of secondary metabolite-related genes (COG-Q) (Fig. 6C). Furthermore, the distribution profiles of most BGC families do not align with the phylogenetic inertia of the core genome tree. For example, several widely distributed BGC families, including RiPPs-125, RiPPs-492, NRPS-364, Terpene-495, and Others-99, are absent in some distant genomes (Fig. 7A), suggesting a convergent loss of secondary metabolism BGCs. We also found that five strains (NCCP_15910, Sample59, AM31-17, ELWA-3_3298, and 7-27-S), which were isolated from human samples, harbored a limited number of BGCs (6, 5, 0, 0, and 0, respectively) (Table S3). Therefore, it can be inferred that an evolutionary strategy in which the loss of secondary metabolite biosynthesis genes may facilitate adaptation to host-associated environments.

3.9. Diverse CAZymes-encoding genes in the paranthracis pan-genome

CAZymes in microbes play pivotal roles in complex carbohydrate metabolism across diverse environments (H. Zhang et al., 2018). In the *B. paranthracis* pan-genome, 8170 CAZyme-coding genes across 177 gene families were identified (Table S6). The majority of these gene families (98, 55.4%) represented the accessory genome, and the remaining 65 (36.7%) and 14 (7.9%) represented the core genome and the strain-specific gene content, respectively (Fig. 7C). The identified CAZymes included auxiliary activities (AAs), carbohydrate-binding molecules (CBMs), carbohydrate esterases (CEs), glycoside hydrolases

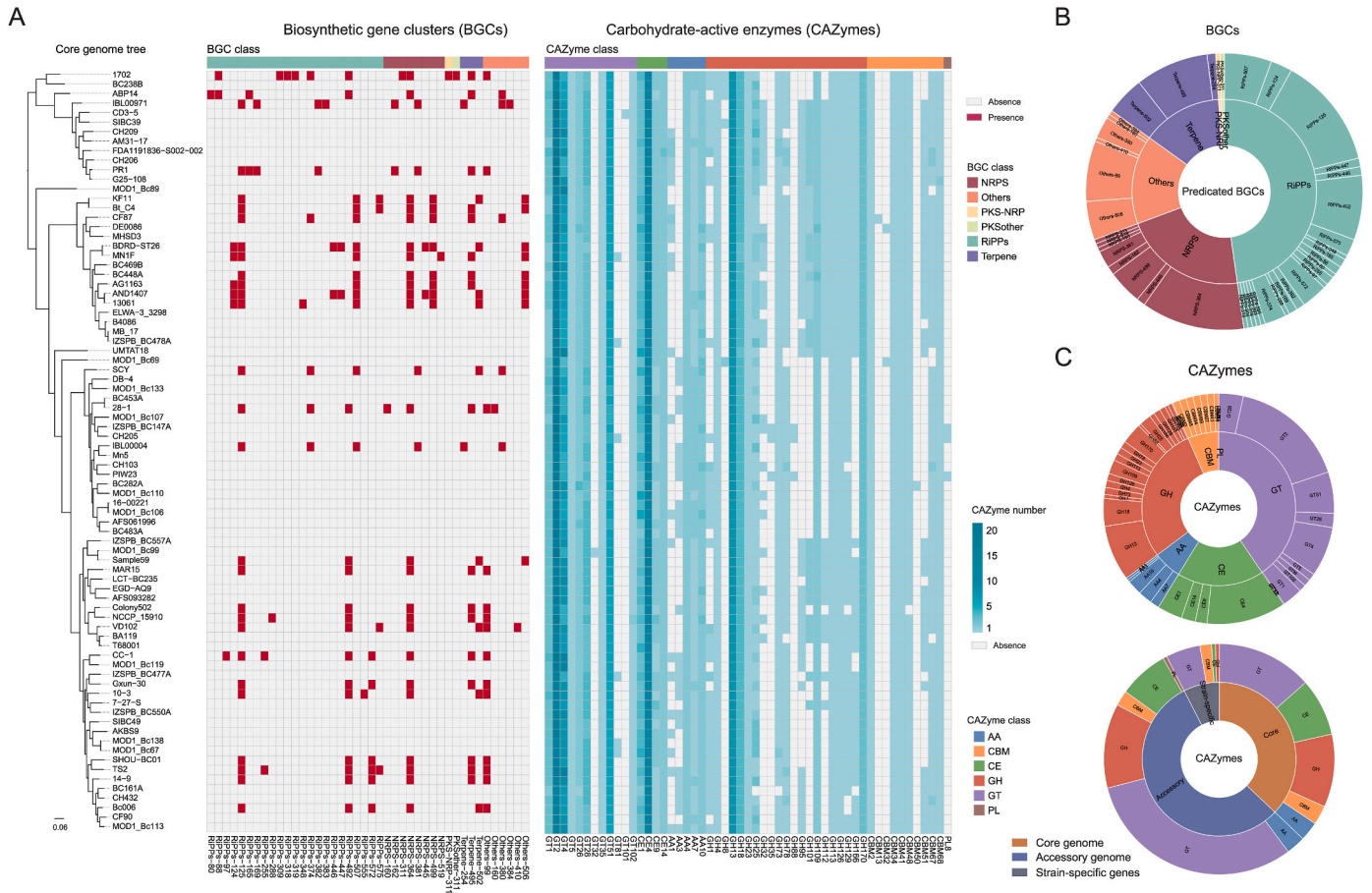

Fig. 7. Secondary metabolite biosynthetic gene clusters (BGCs) and carbohydrate active enzymes (CAZymes) in the *B. paranthracis* pan-genome. A. Heatmap of secondary metabolite BGC and CAZyme distributions. The genome order corresponds to the core genome tree. The presence of BGCs is represented by red blocks, whereas their absence is denoted by white blocks. For CAZymes, the color intensity reflects the gene copy number. B. Sunburst diagram of BGC classification. The inner ring represents classes, and the outer ring represents families classified by BiG-SCAPE. C. Sunburst diagram of CAZyme classification. The class and family categories are represented by inner and outer rings, respectively.

(GHs), glycosyltransferases (GTs), and polysaccharide lyases (PLs), with GTs and GHs being the most abundant classes (Fig. 7C). On average, each genome contained 64.0 ± 0.7 core, 28.4 ± 4.6 accessory, and 0.2 ± 0.5 strain-specific genes encoding 6.1 ± 0.9 AAs, 6.6 ± 0.6 CBMs, 18.7 ± 1.3 CEs, 29.3 ± 2.7 GHs, 40.3 ± 3.1 GTs, and 0.0 ± 0.2 PLs. Among the non-core properties, the majority (75 out of 112) of CAZyme-encoding gene families were present in 10 or fewer strains (Fig. 7A), suggesting that individual variability in metabolic capability may facilitate adaptation to specialized habitats. A total of 91 CAZyme-encoding gene families were inferred to originate from MRCA, 38 of which were affected by gene gain and loss (Fig. S6). This included 26 loss and 22 gain gene families, with the most abundant CAZyme family being GT2. Notably, the GT2 family is involved in the synthesis of bacterial cellulose (Stanisich and Stone, 2009), which constitutes a key component of the extracellular matrix associated with biofilm formation (Chen et al., 2021). Thus, the dynamic changes within the GT2 family appear to reflect adaptive responses of *B. paranthracis*.

Previous studies have revealed that microorganisms have evolved various strategies for secondary metabolite and CAZyme production, enabling them to adapt to their specialized habitats (Otani et al., 2022) (Yin et al., 2022). Moreover, the loss of costly and dispensable CAZyme-encoding genes and secondary metabolite BGCs has been recognized as a survival strategy among bacteria adopting host-associated lifestyles (Ramzi et al., 2019)(Wang et al., 2022). Our findings appear to reflect the special adaptation of *B. paranthracis* to such environments, which aligns with the high relative abundance of *B. paranthracis* in host-associated samples.

3.10. Emerging AMR genes in the paranthracis pan-genome

The *B. cereus* group commonly exhibits resistance to penicillin and other β-lactam antibiotics and has the capacity to develop resistance to frequently used antibiotics, such as ciprofloxacin, cloxacillin, erythromycin, tetracycline, and streptomycin (Citron and Appleman, 2006; Fiedler et al., 2019). *B. paranthracis* isolates are resistant to several antibiotics, including oxacillin, ampicillin, penicillin, and cephalosporin (de Sousa, 2021). In the *B. paranthracis* pan-genome, 39 AMR genes (including allelic variants) were identified, corresponding to resistance to 18 antibiotic classes, including glycopeptide, phenicol, streptogramin, lincosamide, tetracycline, macrolide, and penam (Fig. 8A and B, and Table S7). Four core genes are associated with resistance to fluoroquinolone (*blt*), diaminopyrimidine (*dfrG*), fosfomycin (*fosB*), and rifamycin (*rphB*). Ten of the remaining AMR genes were present in more than half of the *B. paranthracis* genomes, whereas 19 were found in fewer than 10 genomes (Fig. 8A). These non-core AMR genes, which are strain-specific, are likely acquired via mobile units. Plasmids play a key role in microbial ecology and evolution by mediating the horizontal transfer of important genes, especially AMR genes (Andreopoulos et al., 2022). Plasmid nucleotide sequences were detected in 93.4% (226 out of 242) of the *B. paranthracis* genomes (Table S8), indicating that this species is rich in plasmids. Eleven AMR genes, including *aadK*, *bcII*, *fosB*, $vanY_A$, $vanZ_F$, *ugd*, *bcrC*, *Acla_ACT_CHL*, *tet(45)*, and *tet(L)*, were located

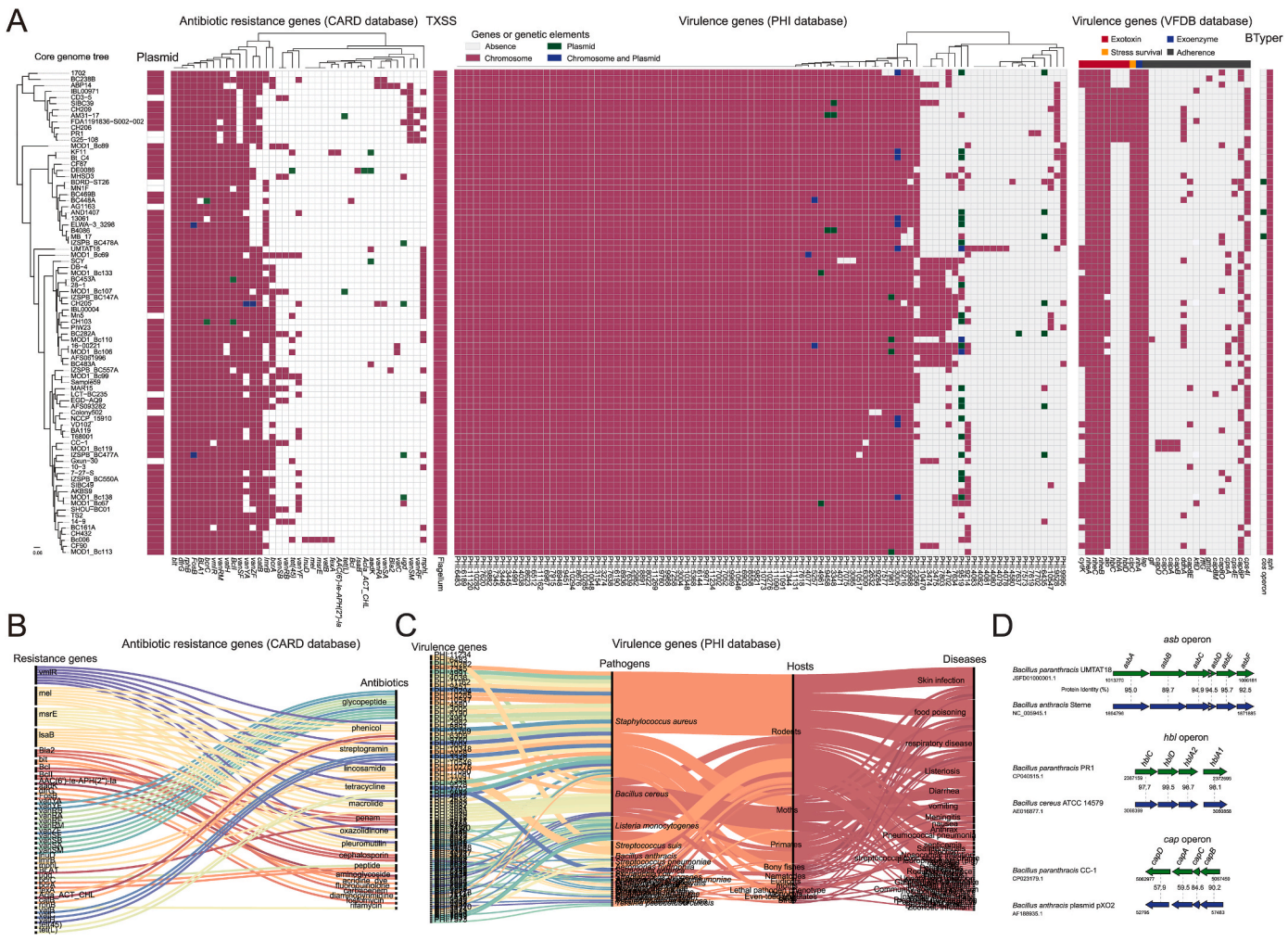

Fig. 8. Distribution of antimicrobial resistance (AMR) and virulence-related genes. A. Heatmap of the distribution of AMR and virulence-related genes. The genome order corresponds to the core genome tree. The presence of these genes is represented by colored blocks, whereas their absence is denoted by white blocks. The detail information of each AMR and virulence-related genes was listed in Table S9 and Table S10. B. Sankey diagram of AMR gene and antibiotic associations. C. Sankey diagram of the interconnections between virulence genes, pathogens, hosts, and diseases. D. Comparison of the genetic organization and protein sequences between homologs of the *asb*, *hbl*, and *cap* operons. The amino acid identities for each pair of homologous proteins are displayed.

on the detected plasmids (Fig. 8A). These genes driven by plasmid-mediated HGT may promote the development of antibiotic resistance in specialized niches of *B. paranthracis*.

In the *B. paranthracis* genome, several *van* genes related to vancomycin resistance are widely present, including *vanR* (80 out of 80), *vanS* (80 out of 80), *vanY* (80 out of 80), and *vanZ* (62 out of 80). However, the absence of *vanH*, *vanA* and *vanX* indicates that the *B. paranthracis* members remain sensitive to vancomycin, as the full complement of seven genes is necessary for vancomycin resistance (Pootoolal et al., 2002). We identified four variant types of *van* genes: *vanA* ($vanR_A$, $vanS_A$, $vanY_A$, and $vanZ_A$), *vanB* ($vanR_B$, $vanS_B$, $vanY_B$, *and* $vanZ_B$), *vanF* ($vanR_F$ and $vanS_F$), and *vanM* ($vanR_M$ and $vanS_M$). These variants have sporadically integrated into the *B. paranthracis* genomes (Fig. 8A), suggesting frequent acquisition of these variants from distant *van* gene clusters. Therefore, emerging vancomycin-resistant *B. paranthracis* strains may arise from the acquisition of *van* genes, which would result in a complete gene cluster. Additionally, *Bacillus* sp. strains isolated from humans or used as human probiotics have been reported to have a significantly high capacity to acquire *van* genes (Sanders et al., 2010) (Cui et al., 2020). The diverse *van* gene profiles within the *B. paranthracis* pan-genome may reflect adaptations to host-associated environments, highlighting the need for continued surveillance and research.

3.11. Virulence-related genetic profile of the paranthracis pan-genome

Within the *B. cereus* group, *B. anthracis*, *B. cereus*, and *B. thuringiensis* are well-studied members, each known for its pathogenicity. To provide a comprehensive view of its pathogenic potential, we investigated the virulence-related genetic profile of *B. paranthracis*. A conserved gene cluster encoding a putative flagellar system was identified in all *B. paranthracis* genomes (Fig. 8A). Flagella in *B. cereus* contribute to motility, adherence, and toxin secretion (Senesi and Ghelardi, 2010) (Enosi Tuipulotu et al., 2021). A total of 96 gene families matching virulence genes listed in the PHI-base database from 24 different pathogens were identified (including *Staphylococcus aureus*, *Bacillus cereus*, *Listeria monocytogenes*, *Streptococcus suis*, and *B. anthracis*) (Table S9 and Fig. 8C). Of these, 72 (73.5%) were present in the majority (more than 70 out of 80) of the *B. paranthracis* genomes. Only one virulence gene family (PHI:2356) was present in more than half of the genomes, whereas 13 were found in fewer than 10 genomes (Fig. 8A). On average, each genome contained 76.2 ± 2.9 potential virulence-related genes. Mutation experiments revealed that the predominant mutant phenotype was "reduced virulence" ($n = 74$, 77.1%) (Table S9), implying that most identified genes were associated with pathogenic potential. The identified virulence genes were associated mainly with various animal hosts, including rodents ($n = 51$), moths ($n = 17$), and primates ($n = 14$), and

were implicated in diseases such as skin infection, food poisoning, respiratory disease, listeriosis, diarrhea, and vomiting (Fig. 8C).

We also identified 27 additional gene families with homology to known virulence genes in the VFDB and BTyper databases (Fig. 8A and Table S10), including important virulence genes responsible for the production of several exotoxins previously reported in *B. cereus* and *B. anthracis*. The *nhe* operon (encoding nonhemolytic enterotoxin), *alo* (encoding anthrolysin O), and *sph* (encoding sphingomyelinase) were present in the core genome, whereas the *cytK* (encoding cytotoxin K), *hbl* (encoding haemolysin BL), and *ces* (encoding cereulide) operons were sporadically distributed, representing accessory gene families. As a potential pathogen, *B. paranthracis* has been reported to cause bacteremia and diarrhea and exhibit cytotoxicity in vitro and non-hemolytic enterotoxicity (22)(23)(45). Given the presence of various virulence-related genes in the *B. paranthracis* pan-genome, the safety of this species for use as probiotic or plant-growth promoting bacteria in biotechnological and agricultural applications requires further evaluation.

3.12. Emerging virulence genes driven by HGT, especially plasmid-mediated HGT

A total of 10 virulence genes were identified on plasmids, with four genes (PHI:3005, PHI:9519, PHI:5257, and PHI:4077) found in both the chromosomes and plasmids of multiple individual genomes (Fig. 8A), suggesting acquisition through plasmid-mediated HGT. Similarly, the *ces* operon was found in the plasmid regions of three distinct genomes. This operon, which is responsible for cereulide synthesis, is located on mega virulence plasmids in the emetic *B. cereus* (Ehling-Schulz et al., 2006). Furthermore, strain UMTAT18 contained a complete petrobactin biosynthesis operon (*asbABCDEF*), which was highly homologous to that of *B. anthracis* Sterne, with 89.7%–95.7% protein sequence identity and identical gene locus organization (Fig. 8D). As an iron-scavenging siderophore, petrobactin is required for growth in macrophages and virulence in mice for *B. anthracis* Sterne (Cendrowski et al., 2004). It also protects cells against oxidative stress and enhances sporulation efficiency in bovine blood (Hagan et al., 2018). Significant homologs and identical genetic organizations were also observed in the *hbl* and *cap* operons (Fig. 8D). The *hbl* operon in *B. cereus* encodes a pore-forming toxin that possesses hemolytic, cytotoxic, dermonecrotic, and vascular permeability activities (Beecher et al., 1995). The *cap* operon located on the *B. anthracis* plasmid encodes membrane associated enzymes that are essential for immune modulation, antiphagocytosis, systemic invasion, and dissemination within the bloodstream (Candela and Fouet, 2005) (Jang et al., 2011). The occurrence of these virulence-related genes/operons likely arises from HGT, especially plasmid-mediated HGT. Given the results of the AMR genes, it can be inferred that plasmids may play a significant role in the evolutionary dynamics of *B. paranthracis*, facilitating pathogenicity and adaptation to specialized host niches. This finding aligns with previous studies indicating that the acquisition of virulence plasmids is an evolutionary trait of many well-known pathogens within the *B. cereus* group, such as *B. anthracis* and *B. thuringiensis* (Méric et al., 2018; Lee et al., 2022). Therefore, our results further support evidence that *B. paranthracis* members may increase pathogenicity and adapt to host niches by acquiring exogenous virulence and AMR genes, thereby posing emerging threats to public health.

4. Conclusion

This study provides comprehensive insights into the biogeographic distribution, pan-genome evolution, and genotypic profiles of key properties of *B. paranthracis*, including secondary metabolism, CAZymes, AMR, and virulence. Metagenomic read recruitment analyses revealed that *B. paranthracis* members are globally distributed, with specific abundances in host-associated samples, indicating that this species is a niche generalist with specialized host adaptation. The open pan-genome, characterized by a flexible gene repertoire in the accessory genome and strain-specific genes, exhibits extensive genetic diversity. Significant differences in functional enrichment and natural selection between the core and accessory genomes indicate distinct evolutionary strategies among pan-genome components. Owing to its flexible function, the accessory genome results from massive adaptive gene losses and gene gains through HGT and has experienced weak purifying selection or positive selection. In contrast, the core genome of *B. paranthracis* is more conserved and has experienced stronger purifying selection, indicating a tendency to preserve essential biological functions. However, we also found that the *B. paranthracis* core genome has experienced gene gains, HGT, and a significant bias toward containing positively selected mutations. Our results suggest that the core genome, which is typically viewed as conserved, also plays a pivotal role in the adaptive evolution of *B. paranthracis*.

Gene loss driving genome reduction represents a predominant evolutionary scenario for *B. paranthracis*. The streamlining strategies, which involve decreasing secondary metabolite BGCs and CAZymes-encoding genes while acquiring AMR and virulence genes, reflect special gene gain and loss patterns that facilitate adaptation to host-associated habitats. This finding is consistent with our observation that *B. paranthracis* has a high relative abundance in host-associated samples. Furthermore, *B. paranthracis* genomes harbor diverse AMR and virulence-related genes, highlighting their pathogenic potential. HGT, especially through plasmids, drives the emergence of antibiotic resistance and pathogenicity within the *B. paranthracis* pan-genome, emphasizing the emerging public health risk of this foodborne pathogen.

CRediT authorship contribution statement

Yuhui Du: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Writing – original draft. **Chengqian Qian:** Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. **Xianxin Li:** Data curation, Formal analysis, Methodology, Visualization. **Xinqian Zheng:** Methodology, Resources, Validation. **Shoucong Huang:** Project administration, Resources, Supervision. **Zhiqiu Yin:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Tingjian Chen:** Funding acquisition, Project administration, Resources, Writing – review & editing. **Li Pan:** Conceptualization, Investigation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by the Science and Technology Project of Guangzhou, China (2023A04J1470), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019ZT08Y318), the Guangdong Provincial Pearl River Talents Program (2019QN01Y228), and the Key Laboratory of Advanced Technology Enterprise of Guangdong Seasoning Food Bio Fermentation (2017B030302002). The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported. Enterprise of Guangdong Seasoning Food Bio Fermentation (2017B030302002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crfs.2024.100867.

References

Abby, S.S., Rocha, E.P.C., 2017. Identification of protein secretion systems in bacterial genomes using MacSyFinder. Methods Mol. Biol. 1615 (October 2015), 1–21. https://doi.org/10.1007/978-1-4939-7033-9_1.

Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.L.V., Cheng, A.A., Liu, S., Min, S.Y., Miroshnichenko, A., Tran, H.K., Werfalli, R.E., Nasir, J.A., Oloni, M., Speicher, D.J., Florescu, A., Singh, B., et al., 2020. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 48 (D1), D517–D525. https://doi.org/10.1093/nar/gkz935.

Anani, H., Zgheib, R., Hasni, I., Raoult, D., Fournier, P.-E., 2020. Interest of bacterial pangenome analyses in clinical microbiology. Microb. Pathog. 149, 104275. https://doi.org/10.1016/j.micpath.2020.104275.

Andreopoulos, W.B., Geller, A.M., Lucke, M., Balewski, J., Clum, A., Ivanova, N.N., Levy, A., 2022. Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. Nucleic Acids Res. 50 (3), e17. https://doi.org/10.1093/nar/gkab1115.

Arnold, B.J., Huang, I.-T., Hanage, W.P., 2022. Horizontal gene transfer and adaptive evolution in bacteria. Nat. Rev. Microbiol. 20 (4), 206–218. https://doi.org/10.1038/s41579-021-00650-4.

Azarian, T., Huang, I.-T., Hanage, W.P., 2020. In: Tettelin, H., Medini, D. (Eds.), Structure and Dynamics of Bacterial Populations: Pangenome Ecology, pp. 115–128. https://doi.org/10.1007/978-3-030-38281-0_5.

Babic, A., Berkmen, M.B., Lee, C.A., Grossman, A.D., 2011. Efficient gene transfer in bacterial cell chains. mBio 2 (2). https://doi.org/10.1128/mBio.00027-11.

Bagci, C., Bryant, D., Cetinkaya, B., Huson, D.H., 2021. Microbial phylogenetic context using phylogenetic outlines. Genome Biology and Evolution 13 (9). https://doi.org/10.1093/gbe/evab213.

Baltrus, D.A., 2013. Exploring the costs of horizontal gene transfer. Trends Ecol. Evol. 28 (8), 489–495. https://doi.org/10.1016/j.tree.2013.04.002.

Beecher, D.J., Schoeni, J.L., Wong, A.C., 1995. Enterotoxic activity of hemolysin BL from Bacillus cereus. Infect. Immun. 63 (11), 4423–4428. https://doi.org/10.1128/iai.63.11.4423-4428.1995.

Bogdanowicz, D., Giaro, K., Wróbel, B., 2012. TreeCmp: comparison of trees in polynomial time. Evol. Bioinf. Online 8, 475–487. https://doi.org/10.4137/EBO.S9657.

Bohlin, J., Eldholm, V., Pettersson, J.H.O., Brynildsrud, O., Snipen, L., 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genom. 18 (1), 151. https://doi.org/10.1186/s12864-017-3543-7.

Bourguignon, T., Kinjo, Y., Villa-Martín, P., Coleman, N.V., Tang, Q., Arab, D.A., Wang, Z., Tokuda, G., Hongoh, Y., Ohkuma, M., Ho, S.Y.W., Pigolotti, S., Lo, N., 2020. Increased mutation rate is linked to genome reduction in prokaryotes. Curr. Biol. : CB 30 (19), 3848–3855.e4. https://doi.org/10.1016/j.cub.2020.07.034.

Boutte, C.C., Crosson, S., 2013. Bacterial lifestyle shapes stringent response activation. Trends Microbiol. 21 (4), 174–180. https://doi.org/10.1016/j.tim.2013.01.002.

Brockhurst, M.A., Harrison, E., Hall, J.P.J., Richards, T., McNally, A., MacLean, C., 2019. The ecology and evolution of pangenomes. Curr. Biol. 29 (20), R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12 (1), 59–60. https://doi.org/10.1038/nmeth.3176.

Bukharin, O.V., Perunova, N.B., Andryuschenko, S.V., Ivanova, E.V., Bondarenko, T.A., Chainikova, I.N., 2019. Genome sequence announcement of Bacillus paranthracis strain ICIS-279, isolated from human intestine. Microbiology Resource Announcements 8 (44). https://doi.org/10.1128/MRA.00662-19.

Candela, T., Fouet, A., 2005. Bacillus anthracis CapD, belonging to the gamma-glutamyltranspeptidase family, is required for the covalent anchoring of capsule to peptidoglycan. Mol. Microbiol. 57 (3), 717–726. https://doi.org/10.1111/j.1365-2958.2005.04718.x.

Carroll, L.M., Wiedmann, M., Mukherjee, M., Nicholas, D.C., Mingle, L.A., Dumas, N.B., Cole, J.A., Kovac, J., 2019. Characterization of emetic and diarrheal Bacillus cereus strains from a 2016 foodborne outbreak using whole-genome sequencing: addressing the microbiological, epidemiological, and bioinformatic challenges. Front. Microbiol. 10, 144. https://doi.org/10.3389/fmicb.2019.00144.

Carroll, L.M., Cheng, R.A., Kovac, J., 2020a. No assembly required: using BTyper3 to assess the congruency of a proposed taxonomic framework for the Bacillus cereus group with historical typing methods. Front. Microbiol. 11, 580691. https://doi.org/10.3389/fmicb.2020.580691.

Carroll, L.M., Wiedmann, M., Kovac, J., 2020b. Proposal of a taxonomic nomenclature for the Bacillus cereus group which reconciles genomic definitions of bacterial species with clinical and industrial phenotypes. mBio 11 (1). https://doi.org/10.1128/mBio.00034-20.

Carroll, L.M., Cheng, R.A., Wiedmann, M., Kovac, J., 2022. Keeping up with the Bacillus cereus group: taxonomy through the genomics era and beyond. Crit. Rev. Food Sci. Nutr. 62 (28), 7677–7702. https://doi.org/10.1080/10408398.2021.1916735.

Castillo-Ramírez, S., Harris, S.R., Holden, M.T.G., He, M., Parkhill, J., Bentley, S.D., Feil, E.J., 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog. 7 (7), e1002129. https://doi.org/10.1371/journal.ppat.1002129.

Cendrowski, S., MacArthur, W., Hanna, P., 2004. Bacillus anthracis requires siderophore biosynthesis for growth in macrophages and mouse virulence. Mol. Microbiol. 51 (2), 407–417. https://doi.org/10.1046/j.1365-2958.2003.03861.x.

Chan, A.P., Sutton, G., DePew, J., Krishnakumar, R., Choi, Y., Huang, X.-Z., Beck, E., Harkins, D.M., Kim, M., Lesho, E.P., Nikolich, M.P., Fouts, D.E., 2015. A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. Genome Biol. 16 (1), 143. https://doi.org/10.1186/s13059-015-0701-6.

Chattopadhyay, S., Weissman, S.J., Minin, V.N., Russo, T.A., Dykhuizen, D.E., Sokurenko, E.V., 2009. High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. Proc. Natl. Acad. Sci. U.S.A. 106 (30), 12412–12417. https://doi.org/10.1073/pnas.0906217106.

Chen, X., Fan, L., Qiu, L., Dong, X., Wang, Q., Hu, G., Meng, S., Li, D., Chen, J., 2021. Metagenomics analysis reveals compositional and functional differences in the gut microbiota of red swamp crayfish, Procambarus clarkii, grown on two different culture environments. Front. Microbiol. 12, 735190. https://doi.org/10.3389/fmicb.2021.735190.

Citron, D.M., Appleman, M.D., 2006. In vitro activities of daptomycin, ciprofloxacin, and other antimicrobial agents against the cells and spores of clinical isolates of Bacillus species. J. Clin. Microbiol. 44 (10), 3814–3818. https://doi.org/10.1128/JCM.00881-06.

Cui, Y., Wang, S., Ding, S., Shen, J., Zhu, K., 2020. Toxins and mobile antimicrobial resistance genes in Bacillus probiotics constitute a potential risk for One Health. J. Hazard Mater. 382, 121266. https://doi.org/10.1016/j.jhazmat.2019.121266.

Cummins, E.A., Hall, R.J., McInerney, J.O., McNally, A., 2022. Prokaryote pangenomes are dynamic entities. Curr. Opin. Microbiol. 66, 73–78. https://doi.org/10.1016/j.mib.2022.01.005.

de Sousa, L.P., 2021. Genomic and pathogenicity of a Bacillus paranthracis isolated from book page surface. Infect. Genet. Evol. : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases 92, 104867. https://doi.org/10.1016/j.meegid.2021.104867.

Didelot, X., Wilson, D.J., 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput. Biol. 11 (2), 1–18. https://doi.org/10.1371/journal.pcbi.1004041.

Du, Y., Zou, J., Yin, Z., Chen, T., 2023. Pan-Chromosome and comparative analysis of agrobacterium fabrum reveal important traits concerning the genetic diversity, evolutionary dynamics, and niche adaptation of the species. Microbiol. Spectr. 11 (2), e0292422. https://doi.org/10.1128/spectrum.02924-22.

Earl, D.A., vonHoldt, B.M., 2012. Structure harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4 (2), 359–361. https://doi.org/10.1007/s12686-011-9548-7.

Ehling-Schulz, M., Fricker, M., Grallert, H., Rieck, P., Wagner, M., Scherer, S., 2006. Cereulide synthetase gene cluster from emetic Bacillus cereus: structure and location on a mega virulence plasmid related to Bacillus anthracis toxin plasmid pXO1. BMC Microbiol. 6, 20. https://doi.org/10.1186/1471-2180-6-20.

Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16 (1), 157. https://doi.org/10.1186/s13059-015-0721-2.

Enosi Tuipulotu, D., Mathur, A., Ngo, C., Man, S.M., 2021. Bacillus cereus: epidemiology, virulence factors, and host-pathogen interactions. Trends Microbiol. 29 (5), 458–471. https://doi.org/10.1016/j.tim.2020.09.003.

Epstein, B., Tiffin, P., 2021. Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes. Proceedings. Biological Sciences 288 (1942), 20201804. https://doi.org/10.1098/rspb.2020.1804.

Everitt, R.G., Didelot, X., Batty, E.M., Miller, R.R., Knox, K., Young, B.C., Bowden, R., Auton, A., Votintseva, A., Larner-Svensson, H., Charlesworth, J., Golubchik, T., Ip, C. L.C., Godwin, H., Fung, R., Peto, T.E.A., Walker, A.S., Crook, D.W., Wilson, D.J., 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nat. Commun. 5, 3956. https://doi.org/10.1038/ncomms4956.

Falush, D., Stephens, M., Pritchard, J.K., 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol. Ecol. Notes 7 (4), 574–578. https://doi.org/10.1111/j.1471-8286.2007.01758.x.

Fiedler, G., Schneider, C., Igbinosa, E.O., Kabisch, J., Brinks, E., Becker, B., Stoll, D.A., Cho, G.-S., Huch, M., Franz, C.M.A.P., 2019. Antibiotics resistance and toxin profiles of Bacillus cereus-group isolates from fresh vegetables from German retail markets. BMC Microbiol. 19 (1), 250. https://doi.org/10.1186/s12866-019-1632-2.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39 (Web Server issue), W29–W37. https://doi.org/10.1093/nar/gkr367.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2015. Expanded Microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43 (D1), D261–D269. https://doi.org/10.1093/nar/gku1223.

Goyal, A., 2018. Metabolic adaptations underlying genome flexibility in prokaryotes. PLoS Genet. 14 (10), e1007763. https://doi.org/10.1371/journal.pgen.1007763.

Hagan, A.K., Plotnick, Y.M., Dingle, R.E., Mendel, Z.I., Cendrowski, S.R., Sherman, D.H., Tripathi, A., Hanna, P.C., 2018. Petrobactin protects against oxidative stress and enhances sporulation efficiency in Bacillus anthracis Sterne. mBio 9 (6). https://doi.org/10.1128/mBio.02079-18.

Hao, W., Golding, G.B., 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16 (5), 636–643. https://doi.org/10.1101/gr.4746406.

Hartmann, F.E., Croll, D., 2017. Distinct trajectories of massive recent gene gains and losses in populations of a microbial eukaryotic pathogen. Mol. Biol. Evol. 34 (11), 2808–2822. https://doi.org/10.1093/molbev/msx208.
Heaps, H.S., 1978. Information Retrieval-Computational and Theoretical Aspects. Academic Press, New York, NY.
Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., Bork, P., 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol. Biol. Evol. 34 (8), 2115–2122. https://doi.org/10.1093/molbev/msx148.
Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61 (6), 1061–1067. https://doi.org/10.1093/sysbio/sys062.
Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9 (1), 5114. https://doi.org/10.1038/s41467-018-07641-9.
Jang, J., Cho, M., Chun, J.-H., Cho, M.-H., Park, J., Oh, H.-B., Yoo, C.-K., Rhie, G., 2011. The poly-γ-D-glutamic acid capsule of Bacillus anthracis enhances lethal toxin activity. Infect. Immun. 79 (9), 3846–3854. https://doi.org/10.1128/IAI.01145-10.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7 : improvements in performance and usability article fast track. Molecular Biology and Evolutionl 30 (4), 772–780. https://doi.org/10.1093/molbev/mst010.
Konstantinidis, K.T., Tiedje, J.M., 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 102 (7), 2567–2572. https://doi.org/10.1073/pnas.0409727102.
Koonin, E.V., Wolf, Y.I., 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36 (21), 6688–6719. https://doi.org/10.1093/nar/gkn668.
Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D., Clavel, T., 2016. IMNGS: a comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Sci. Rep. 6, 33721. https://doi.org/10.1038/srep33721.
Lee, I.P.A., Eldakar, O.T., Gogarten, J.P., Andam, C.P., 2022. Bacterial cooperation through horizontal gene transfer. Trends Ecol. Evol. 37 (3), 223–232. https://doi.org/10.1016/j.tree.2021.11.006.
Li, T., Yin, Y., 2022. Critical assessment of pan-genomic analysis of metagenome-assembled genomes. Briefings Bioinf. 23 (6). https://doi.org/10.1093/bib/bbac413.
Liu, Y., Du, J., Lai, Q., Zeng, R., Ye, D., Xu, J., Shao, Z., 2017. Proposal of nine novel species of the Bacillus cereus group. Int. J. Syst. Evol. Microbiol. 67 (8), 2499–2508. https://doi.org/10.1099/ijsem.0.001821.
Liu, B., Zheng, D., Zhou, S., Chen, L., Yang, J., 2022. Vfdb 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 50 (D1), D912–D917. https://doi.org/10.1093/nar/gkab1107.
Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., Rodrigues, J.F.M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., Sunagawa, S., Bork, P., 2020. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. 14 (5), 1247–1259. https://doi.org/10.1038/s41396-020-0600-z.
Malard, L.A., Anwar, M.Z., Jacobsen, C.S., Pearce, D.A., 2019. Biogeographical patterns in soil bacterial communities across the Arctic region. FEMS (Fed. Eur. Microbiol. Soc.) Microbiol. Ecol. 95 (9). https://doi.org/10.1093/femsec/fiz128.
Matson, M.J., Anzick, S.L., Feldmann, F., Martens, C.A., Drake, S.K., Feldmann, H., Massaquoi, M., Chertow, D.S., Munster, V.J., 2020. Bacillus paranthracis isolate from blood of fatal ebola virus disease case. Pathogens 9 (Issue 6). https://doi.org/10.3390/pathogens9060475.
Medema, M.H., Blin, K., Cimermancic, P., De Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R., 2011. AntiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39 (Suppl. 2), 339–346. https://doi.org/10.1093/nar/gkr466.
Mendes, F.K., Vanderpool, D., Fulton, B., Hahn, M.W., 2021. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 36 (22–23), 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.
Méric, G., Mageiros, L., Pascoe, B., Woodcock, D.J., Mourkas, E., Lamble, S., Bowden, R., Jolley, K.A., Raymond, B., Sheppard, S.K., 2018. Lineage-specific plasmid acquisition and the evolution of specialized pathogens in Bacillus thuringiensis and the Bacillus cereus group. Mol. Ecol. 27 (7), 1524–1540. https://doi.org/10.1111/mec.14546.
Merkl, R., 2006. A comparative categorization of protein function encoded in bacterial or archeal genomic islands. J. Mol. Evol. 62 (1), 1–14. https://doi.org/10.1007/s00239-004-0311-5.
Morales, M., Sentchilo, V., Carraro, N., Causevic, S., Vuarambon, D., van der Meer, J.R., 2023. Fitness-conditional genes for soil adaptation in the bioaugmentation agent Pseudomonas veronii 1YdBTEX2. mSystems 8 (2), e0117422. https://doi.org/10.1128/msystems.01174-22.
Morris, J.J., Lenski, R.E., Zinser, E.R., 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. mBio 3 (2). https://doi.org/10.1128/mBio.00036-12.
Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., Scheffler, K., 2013. FUBAR: a fast, unconstrained bayesian AppRoximation for inferring selection. Mol. Biol. Evol. 30 (5), 1196–1205. https://doi.org/10.1093/molbev/mst030.
Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L.T.D., Goering, A.W., Thomson, R.J., Metcalf, W.W., Kelleher, N.L., Barona-Gomez, F., Medema, M.H., 2020. A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. 16 (1), 60–68. https://doi.org/10.1038/s41589-019-0400-9.
Otani, H., Udwary, D.W., Mouncey, N.J., 2022. Comparative and pangenomic analysis of the genus Streptomyces. Sci. Rep. 12 (1), 18909. https://doi.org/10.1038/s41598-022-21731-1.
Pande, S., Merker, H., Bohl, K., Reichelt, M., Schuster, S., de Figueiredo, L.F., Kaleta, C., Kost, C., 2014. Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. ISME J. 8 (5), 953–962. https://doi.org/10.1038/ismej.2013.211.
Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25 (7), 1043–1055. https://doi.org/10.1101/gr.186072.114.
Pootoolal, J., Neu, J., Wright, G.D., 2002. Glycopeptide antibiotic resistance. Annu. Rev. Pharmacol. Toxicol. 42, 381–408. https://doi.org/10.1146/annurev.pharmtox.42.091601.142813.
Power, J.J., Pinheiro, F., Pompei, S., Kovacova, V., Yüksel, M., Rathmann, I., Förster, M., Lässig, M., Maier, B., 2021. Adaptive evolution of hybrid bacteria by horizontal gene transfer. Proc. Natl. Acad. Sci. U.S.A. 118 (10). https://doi.org/10.1073/pnas.2007873118.
Preska Steinberg, A., Lin, M., Kussell, E., 2022. Core genes can have higher recombination rates than accessory genes within global microbial populations. Elife 11, e78533. https://doi.org/10.7554/eLife.78533.
Ramzi, A.B., Che Me, M.L., Ruslan, U.S., Baharum, S.N., Nor Muhammad, N.A., 2019. Insight into plant cell wall degradation and pathogenesis of Ganoderma boninense via comparative genome analysis. PeerJ 7, e8065. https://doi.org/10.7717/peerj.8065.
Rasigade, J.-P., Hollandt, F., Wirth, T., 2018. Genes under positive selection in the core genome of pathogenic Bacillus cereus group members. Infect. Genet. Evol. : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases 65, 55–64. https://doi.org/10.1016/j.meegid.2018.07.009.
Raymond, B., Bonsall, M.B., 2013. Cooperation and the evolutionary ecology of bacterial virulence: the Bacillus cereus group as a novel study system. Bioessays : News and Reviews in Molecular, Cellular and Developmental Biology 35 (8), 706–716. https://doi.org/10.1002/bies.201300028.
Sanders, M.E., Akkermans, L.M.A., Haller, D., Hammerman, C., Heimbach, J., Hörmannsperger, G., Huys, G., Levy, D.D., Lutgendorff, F., Mack, D., Phothirath, P., Solano-Aguilar, G., Vaughan, E., 2010. Safety assessment of probiotics for human use. Gut Microb. 1 (3), 164–185. https://doi.org/10.4161/gmic.1.3.12127.
Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
Senesi, S., Ghelardi, E., 2010. Production, secretion and biological activity of Bacillus cereus enterotoxins. Toxins 2 (7), 1690–1703. https://doi.org/10.3390/toxins2071690.
Seshadri, R., Roux, S., Huber, K.J., Wu, D., Yu, S., Udwary, D., Call, L., Nayfach, S., Hahnke, R.L., Pukall, R., White, J.R., Varghese, N.J., Webb, C., Palaniappan, K., Reimer, L.C., Sardà, J., Bertsch, J., Mukherjee, S., Reddy, T.B.K., et al., 2022. Expanding the genomic encyclopedia of Actinobacteria with 824 isolate reference genomes. Cell Genomics 2 (12), 100213. https://doi.org/10.1016/j.xgen.2022.100213.
Sharrar, A.M., Crits-Christoph, A., Méheust, R., Diamond, S., Starr, E.P., Banfield, J.F., 2020. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. mBio 11 (3), e00416. https://doi.org/10.1128/mBio.00416-20, 20.
Shen, Y., Yang, H., Lin, Z., Chu, L., Pan, X., Wang, Y., Liu, W., Jin, P., Miao, W., 2023. Screening of compound-formulated Bacillus and its effect on plant growth promotion. Front. Plant Sci. 14, 1174583. https://doi.org/10.3389/fpls.2023.1174583.
Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.a., Alm, E.J., 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480, 241–244. https://doi.org/10.1038/nature10571.
Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., Rubin, E.M., 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318 (5855), 1449–1452. https://doi.org/10.1126/science.1147112.
Stanisich, V.A., Stone, B.A., 2009. *Enzymology and molecular genetics of biosynthetic enzymes for (1,3)-B-glucans: prokaryotes.* Chemistry, Biochemistry. Biology of 1-3 Beta Glucans and Related Polysaccharides.
Szabová, J., Ruzicka, P., Verner, Z., Hampl, V., Lukes, J., 2011. Experimental examination of EFL and MATX eukaryotic horizontal gene transfers: coexistence of mutually exclusive transcripts predates functional rescue. Mol. Biol. Evol. 28 (8), 2371–2378. https://doi.org/10.1093/molbev/msr060.
Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P.H., Godelle, B., 1997. Role of mutator alleles in adaptive evolution. Nature 387 (6634), 700–702. https://doi.org/10.1038/42696.
Tamura, K., Stecher, G., Kumar, S., 2021. MEGA11: molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. 38 (7), 3022–3027. https://doi.org/10.1093/molbev/msab120.
Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J. P., Nelson, W.C., et al., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U.S.A. 102 (39), 13950–13955. https://doi.org/10.1073/pnas.0506758102.

Tettelin, H., Riley, D., Cattuto, C., D, M., 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. 11 (5), 472–477. https://doi.org/10.1016/j.mib.2008.09.006.

Thomas, T., Moitinho-Silva, L., Lurgi, M., Björk, J.R., Easson, C., Astudillo-García, C., Olson, J.B., Erwin, P.M., López-Legentil, S., Luter, H., Chaves-Fonnegra, A., Costa, R., Schupp, P.J., Steindler, L., Erpenbeck, D., Gilbert, J., Knight, R., Ackermann, G., Victor Lopez, J., et al., 2016. Diversity, structure and convergent evolution of the global sponge microbiome. Nat. Commun. 7, 11870. https://doi.org/10.1038/ncomms11870.

Touchon, M., Rocha, E.P.C., Abby, S.S., Ne, B., 2014. MacSyFinder : a program to mine genomes for molecular systems with an application to CRISPR-cas systems. PLoS One 9 (10), 1–9. https://doi.org/10.1371/journal.pone.0110726.

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S.Y., De Silva, N., Martinez, M.C., Pedro, H., Yates, A.D., Hassani-Pak, K., Hammond-Kosack, K.E., 2020. PHI-base: the pathogen-host interactions database. Nucleic Acids Res. 48 (D1), D613–D620. https://doi.org/10.1093/nar/gkz904.

von Meijenfeldt, F.A.B., Hogeweg, P., Dutilh, B.E., 2023. A social niche breadth score reveals niche range strategies of generalists and specialists. Nature Ecology & Evolution 7 (5), 768–781. https://doi.org/10.1038/s41559-023-02027-7.

Wang, Y., Wu, J., Yan, J., Guo, M., Xu, L., Hou, L., Zou, Q., 2022. Comparative genome analysis of plant ascomycete fungal pathogens with different lifestyles reveals distinctive virulence strategies. BMC Genom. 23 (1), 34. https://doi.org/10.1186/s12864-021-08165-1.

Wu, H., Wang, D., Gao, F., 2021. Toward a high-quality pan-genome landscape of Bacillus subtilis by removal of confounding strains. Briefings Bioinf. 22 (2), 1951–1971. https://doi.org/10.1093/bib/bbaa013.

Wu, H., Yang, Z.-K., Yang, T., Wang, D., Luo, H., Gao, F., 2022. An effective preprocessing method for high-quality pan-genome analysis of Bacillus subtilis and Escherichia coli. Methods Mol. Biol. 2377, 371–390. https://doi.org/10.1007/978-1-0716-1720-5_21.

Yang, T., Gao, F., 2022. High-quality pan-genome of Escherichia coli generated by excluding confounding and highly similar strains reveals an association between unique gene clusters and genomic islands. Briefings Bioinf. 23 (4). https://doi.org/10.1093/bib/bbac283.

Yin, Z., Wang, X., Hu, Y., Zhang, J., Li, H., Cui, Y., Zhao, D., Dong, X., Zhang, X., Liu, K., Du, B., Ding, Y., Wang, C., 2022. Metabacillus dongyingensis sp. nov. Is represented by the plant growth-promoting bacterium BY2G20 isolated from saline-alkaline soil and enhances the growth of Zea mays L. Under salt stress. mSystems 8 (1), e0142621. https://doi.org/10.1128/msystems.01426-21.

Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., Chun, J., 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int. J. Syst. Evol. Microbiol. 67 (5), 1613–1617. https://doi.org/10.1099/ijsem.0.001755.

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., Dai, L., 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem. Biophys. Res. Commun. 419 (4), 779–781. https://doi.org/10.1016/j.bbrc.2012.02.101.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., Yin, Y., 2018. DbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 46 (W1), W95–W101. https://doi.org/10.1093/nar/gky418.

Zhong, C., Han, M., Yu, S., Yang, P., Li, H., Ning, K., 2018. Pan-genome analyses of 24 Shewanella strains re-emphasize the diversification of their functions yet evolutionary dynamics of metal-reducing pathway. Biotechnol. Biofuels 11 (1), 1–13. https://doi.org/10.1186/s13068-018-1201-1.

Zhong, Chaofang, Han, Maozhen, Yang, Pengshuo, Chen, Chaoyun, Yu, Hui, Lusheng Wang, K.N., 2019. Comprehensive analysis reveals the evolution and pathogenicity of aeromonas, viewed from both single isolated species and microbial communities. mSystems 4 (5), e00252, 19.

Zhu, Q., Kosoy, M., Dittmar, K., 2014. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. BMC Genom. 15 (1), 717. https://doi.org/10.1186/1471-2164-15-717.